SICS Technical Report
T2009:09
ISSN : 1100-3154

Information Access in a Multilingual World: Transitioning from Research to Real-World Applications

by

Frederic Gey , Jussi Karlgren and Noriko Kando

Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, SWEDEN

Information Access in a Multilingual World: Transitioning from Research to Real-World Applications

Fredric Gey, University of California, Berkeley
Jussi Karlgren, Swedish Institute of Computer Science, Stockholm
Noriko Kando, National Institute of Informatics, Tokyo

July 23, 2009

Abstract

This report constitutes the proceedings of the workshop on *Information Access in a Multilingual World: Transitioning from Research to Real-World Applications*, held at SIGIR 2009 in Boston, July 23, 2009.

Multilingual Information Access (MLIA) is at a turning point wherein substantial real-world applications are being introduced after fifteen years of research into cross-language information retrieval, question answering, statistical machine translation and named entity recognition. Previous workshops on this topic have focused on research and small-scale applications. The focus of this workshop was on technology transfer from research to applications and on what future research needs to be done which facilitates MLIA in an increasingly connected multilingual world.

SICS Technical Report T2009:09
ISRN: SICS-T–2009/09-SE
ISSN: 1100-3154

Papers and talks presented at the workshop

Introduction by Frederic Gey, Jussi Karlgren, and Noriko Kando
(Also published in SIGIR Forum, December 2009, Volume 43 Number 2, pp. 24-28)

Keynote talk by Ralf Steinberger
Presenting the Joint Research Centre of the European Commission's multilingual media monitoring and analysis applications, including NewsExplorer (http://press.jrc.it/overview.html)

Fredric Gey
Romanization – An Untapped Resource for Out-of-Vocabulary Machine Translation for CLIR

John I. Tait
What's wrong with Cross-Lingual IR?

David Nettleton, Mari-Carmen Marcos, Bartolom Mesa
User Study of the Assignment of Objective and Subjective Type Tags to Images in Internet, considering Native and non Native English Language Taggers

Elena Filatova
Multilingual Wikipedia, Summarization, and Information Trustworthiness

Michael Yoshitaka Erlewine
Ubiquity: Designing a Multilingual Natural Language Interface

Masaharu Yoshioka
NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites

Elena Montiel-Ponsoda, Mauricio Espinoza, Guadalupe Aguado de Cea
Multilingual Ontologies for Information Access

Jiangping Chen, Miguel Ruiz
Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries

Wei Che (Darren) Huang, Andrew Trotman, Shlomo Geva
A Virtual Evaluation Track for Cross Language Link Discovery

Kashif Riaz
Urdu is not Hindi for Information Access

Hideki Isozaki, Tsutomu Hirao, Katsuhito Sudoh, Jun Suzuki, Akinori Fujino, Hajime Tsukada, Masaaki Nagata
A Patient Support System based on Crosslingual IR and Semi-supervised Learning

1 Introduction and Overview

The workshop *Information Access in a Multilingual World: Transitioning from Research to Real-World Applications* was held at SIGIR 2009 in Boston, July 23, 2009. The workshop was held in cooperation with the InfoPlosion Project of Japan[1] The workshop was the third workshop on the topic of multilingual information access held at SIGIR conferences this decade. The first, at SIGIR 2002 in Tampere, was on the topic of "Cross Language Information Retrieval: A Research Roadmap". The second was at SIGIR 2006 on the topic of "New Directions in Multilingual Information Access". Over the past decade the field has matured and significant real world applications have appeared. Our goal in this 2009 workshop was to collate experiences and plans for the real-world application of multilingual technology to information access. Our aim was to identify the remaining barriers to practical multilingual information access, both technological and from the point of view of user interaction. We were fortunate to obtain as invited keynote speaker Dr Ralf Steinberger of the Joint Research Centre (JRC) of the European Commission, presenting the Joint Research Centre's multilingual media monitoring and analysis applications, including NewsExplorer. Dr. Steinberger provided an overview paper about their family of applications, which was the first paper in the workshop proceedings.

In our call for papers we specified two types of papers, research papers and position papers. Of the 15 papers initially submitted, two were withdrawn and two were rejected. We accepted 3 research papers and 8 position papers, covering topics from evaluation (of image indexing and of cross-language information retrieval in general), Wikipedia and trust, news site characterization, multilinguality in digital libraries, multilingual user interface design, access to less commonly taught languages (e.g. Indian subcontinent languages), implementation and application to health care. We feel these papers represent a cross-section of the work remaining to be done in moving toward full information access in a multilingual world.

2 Keynote Address

The opening session was the keynote address on "Europe Media Monitoring Family of Applications." Dr. Ralf Steinberger presented a detailed overview of a major initiative of the European Commission's Joint Research Center at Ispra, Italy to provide just-in-time access to large scale worldwide news feeds in approximately 50 languages. At the heart of the system is the Europe Media Monitor news data acquisition from about 2,200 web news sources to gather between 80,000 and 100,000 news articles daily (on average). The 'monitor' visits news web sites up to every five minutes for latest news articles. The news gathering engine feeds its articles into four public news analysis systems:

NewsBrief – which provides real-time (every ten minutes) news clustering and classification, breaking news detection, and an email subscription

[1] http://www.infoplosion.nii.ac.jp/info-plosion/ctr.php/m/IndexEng/a/Index/

facility MedISys – a real-time system which filters out only news reports of a public health nature, including threats of chemical, biological, radiological and nuclear nature NewsExplorer – which displays a daily clustered view of the major news items for each of the 19 languages covered, performs a long-term trend analysis, and offers entity pages showing information gathered in the course of years for each entity, including person titles, multilingual name variants, reported speech quotations, and relations. Languages cover 14 European Union languages plus Arabic, Farsi, Norwegian, Russian, and Turkish. EMM-Labs – which includes a suite of tools for media-focused text mining and visualization, including various map representation of the news, multilingual event extraction, and social network browsers.

3 Research Papers

The research paper by Nettleton, Marcos, and Mesa-Lao of Barcelona, Spain, "The Assignment of Tags to Images in Internet: Language Skill Evaluation" was presented by Ricardo Baeza-Yates. The authors had performed a study on differences between native and non-native users when labeling images with verbal tags. One of the results presented was that the diversity was lower for non-native users, reasonably explained through their relatively smaller vocabulary. The authors studied tags related to concrete image characteristics separately from tags related to emotions evoked by the image: they found, again reasonable in view of likely relative exposure of users to concrete and abstract terminology, that the difference was greater for evocative terms than for concrete visual terms. This study elegantly demonstrated the limits of linguistic competence between native and non-native, simultaneously giving rise to discussion of which usage is the more desirable in a tagging application: do we really wish to afford users the full freedom to choose any term, when many users are likely to be content with a more constrained variation in terminology?

Elena Filatova of Fordham University USA, presented her paper on "Multilingual Wikipedia, Summarization, and Information Trustworthiness." Her experiment showed how a multilingual resource such as Wikipedia can be leveraged to serve as a summarization tool: sentences were matched across languages using an established algorithm to find similarities across languages. Sentences that were represented in many languages were judged as more useful for the purposes of the summary than others. This judgment was verified by having readers assess the quality of summaries. The research corpus was a subset of Wikipedia on biographies utilized in the DUC (Document Understanding Conference) 2004 evaluation.

The paper "A Virtual Evaluation Track for Cross Language Link Discovery" by Huang, Trotman and Geva was presented by Shlomo Geva of Queensland University of Technology, Australia. The authors propose a new evaluation shared task for INEX, NTCIR and CLEF, where participating projects will contribute towards an interlinked universe of shared information across languages, based on internet materials. The objective is to create a low-footprint evaluation campaign, which can be performed off-line, asynchronously, and in a distributed fashion.

4 Position Papers

Masaharu Yoshioka of Hokkaido University, Japan presented a paper on "NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites" Yoshioka's idea was that news sites from different countries in different languages might provide unique viewpoints of reporting the same news stories. The NSContrast system uses "contrast set mining (which) aims to extract the characteristic information about each news site by performing term co-occurrence analysis." To test the ideas, a news article database was assembled from China, Japan, Korea and the USA (representing the 4 languages of these countries). In order to compensate for poor or missing translation, Wikipedia in these languages was mined for named entity translation equivalents.

John Tait of the Information Retrieval Facility in Vienna, Austria, presented a provocative view of "What's wrong with Cross-Lingual IR?" Tait argued that laboratory-based evaluations as found in TREC and other evaluation campaigns have limited generalizability to large scale real-world application venues. In particular, patent searches within the patent intellectual property domain involve a complex and iterative process. Searches have a heavy recall emphasis to validate (or invalidate) patent applications. Moreover, in order to validate the novelty of a patent application, patents in any language must be searched, but the current dominance is with English, Japanese, and possibly Korean. In the future, Chinese will become a major patent language for search focus.

Jiangpen Chen presented her paper co-authored with Miguel Ruiz "Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries." The paper described a range of services which are and might be provided by digital libraries, including multilingual information access. The authors described an integrative cross-lingual information access framework in which cross-language search was supplemented by translational knowledge which integrates different resources to develop a lexical knowledge base by enlisting, among other, the users of the systems to participate in the development of the system capability. Chen's presentation provided a number of example systems which provided some level of bilingual capability upon which future systems might be modeled.

Michael Yoshitaka Erlewine of Mozilla Labs (now at MIT in Linguistics) presented a paper "Ubiquity: Designing a Multilingual Natural Language Interface" about the development of a multilingual textual interface for the Firefox browser which aims at an internationalizable natural language interface which aligns with each "user's natural intuitions about their own language's syntax." The shared vision is that we can put theoretical linguistic insights into practice in creating a user interface (and underlying search and browse capability) that provides a universal language parser with minimal settings for a particular language.

Fredric Gey of the University of California, Berkeley (one of the workshop organizers) presented a paper on "Romanization – An Untapped Resource for Out-of-Vocabulary Machine Translation for CLIR." The paper noted that rule-based transliteration (Romanization) of non-European scripts has been devised for over 55 languages by the USA Library of Congress for cataloging books written in non-latin scripts, including many

variations of Cyrillic and the Devanagiri scripts of most Indian sub-continent languages. The paper argued that rule-based Romanization could be combined with approximate string matching to provide cross-lingual named entity recognition for borrowed words (names) which have not yet made it into general bilingual dictionaries or machine-translation software resources. The approach should be especially beneficial for less resourced languages for which parallel corpora are unavailable.

Kashif Riaz of the University of Minnesota presented a paper "Urdu is not Hindi for Information Access." The paper argued for separate research and development for the Urdu language instead of piggy-backing on tools developed for the Hindi language. Urdu, the national language of Pakistan, and Hindi, the major national language of India, share a major common spoken vocabulary such that speakers of each language can be as well-understood by speakers of the other language as if they were dialects of a common language – however written Urdu is represented by the Arabic script while written Hindi is represented by a Devanagari script. The paper differentiates the separate cultural heritage of each language and argues for significant additional and independent natural language processing development for the Urdu language.

The paper "A Patient Support System based on Crosslingual IR and Semi-supervised Learning" by Isozaki and others of NTT Communication Science Laboratories Kyoto, Japan, was presented by Hideki Isozaki. The authors are constructing a system for aiding medical patients in their quest for information concerning their condition, including treatments, medications and trends in treatments. Because considerable medical information is available in English, the system incorporates a cross-language retrieval module from Japanese to English. The content being accessed is both technical articles (PubMed) and patient-run web, government information sites focused on medical conditions and local information about doctors and surgeons. For technical terms which may not be understood or used by patients, the system provides a synonym generator from lay terms to medical terminology. The system's cross-language goal is to analyze multiple English medical documents "with information extraction/data mining technologies" to generate a Japanese survey summarizing the analysis. Currently the system supports medical literature searches (which have high credibility) and is in the process of expanding to patient sites for which credibility judgment criteria and methods will need to be developed.

5 Discussion of the Future of Multilingual Information Access

The final session was a free-ranging discussion of future research needs and the remaining barriers to widespread adoption of well-researched techniques in multilingual information access into real-world applications.

Discussion on what usage needs to be supported by future systems for cross-lingual information access took as its starting point the question of what usage scenarios specifically need technical support. The require-

ments for professional information analysts with a working knowledge of several languages are different from the needs of lay users with no or little knowledge of any second language beyond their own and with only passing knowledge of the task under consideration. Most of the projects presented here did not explicitly address use cases, nor did they formulate any specific scenario of use, other than through implicit design. The long time failure of machine translation systems was mentioned as a negative example: engineering efforts were directed towards the goal of fluent, high quality sentence-by-sentence translation which in fact seldom has been a bottleneck for human language users. The alternative view, held by many, is that most users have been satisfied by approximate translations which convey the content of the original document.

The suggestion was put forth that the field of cross-lingual information access might be best served by a somewhat more systematic approach to modelling the client they are building the system for; that would in turn better inform the technology under consideration and allow system building project to share resources and evaluation mechanisms.

Action items suggested were, among others, creation of a permanent web site dedicated to research and development of multilingual information access. The first task of the web site would be to accumulate and identify available multilingual corpora to be widely distributed as a goal of further development of equal access to information regardless of language.

6 Conclusion

This workshop recognized that the time has come for the significant body of research on cross-language retrieval, translation and named entity recognition to be incorporated into working systems which are scalable and serve real customers. Two example systems were presented, news summarization (by the keynote speaker) and by researchers trying to provide information support for medical patients. In addition another speaker provided an architecture for integrating multilingual information access within the digital library environment, and one presentation suggested a distributed, low-footprint shared task for evaluation purposes. The discussion sessions generated directions and suggested next steps toward this agenda of developing real-world application systems.

These next steps necessarily will involve sharing experiences of real-world deployment and usage across systems and projects. To best encourage and accommodate such joint efforts, those experiences must be documented, published, and presented in some common forum. If evaluation is to proceed beyond system benchmarking, finding and leveraging these common real-world experiences are crucial to achieve valid and sustainable progress for future projects.

Romanization – An Untapped Resource for Out-of-Vocabulary Machine Translation for CLIR

Fredric Gey
University of California, Berkeley
UC Data Archive & Technical Assistance
Berkeley, CA 94720-5100
510-643-1298

gey@berkeley.edu

ABSTRACT

In Cross-Language Information Retrieval (CLIR), the most continuing problem in query translation is the occurrence of out-of-vocabulary (OOV) terms which are not found in the resources available for machine translation (MT), e.g dictionaries, etc. This usually occurs when new named entities appear in news or other articles which have not been entered into the resource. Often these named entities have been phonetically rendered into the target language, usually from English. Phonetic back-transliteration can be achieved in a number of ways. One of these, which has been under-utilized for MT is Romanization, or rule-based transliteration of foreign typescript into the Latin alphabet. We argue that Romanization, coupled with approximate string matching, can become a new resource for approaching the OOV problem

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, linguistic processing*

General Terms

Experimentation

Keywords

Machine Translation, Romanization Cross-Language Information Retrieval

1. INTRODUCTION

Successful cross-language information retrieval requires, at a minimum, the query (or document) in one language be translated correctly into the other language. This may be done using formal bilingual dictionaries or bilingual lexicons created statistically from aligned parallel corpora. But sometimes these resources have limited coverage with respect to current events, especially named entities such as new people or obscure places have appeared in news stories and their translation has yet to emerge within parallel corpora or enter into formal dictionaries. In addition, a plethora of name variants also confuse the issue of named entity recognition. Steinberger and Pouliquen (2007) discuss these issues in detail when dealing with multilingual news summarization. For non-Latin scripts, this becomes particularly problematic because the user of western scripted languages (such as in USA, England, and most of Europe) cannot guess phonetically what the name might be in his/her native language, even if the word or phrase was borrowed from English in the first place. In many cases, borrowed words enter the language as a phonetic rendering, or transliteration or the original language word. For example, the Japanese word **コンピュータ** (computer). Knight and Graehl (1997) jump-started transliteration research, particularly for Japanese-English by developing a finite state machine for phonetic recognition between the two languages. The phonetic transliteration of the above Japanese is 'konpyuutaa'.

There is, however, an alternative to phonetic transliteration, and that is Romanization, or rule-based rendering of a foreign script into the Latin alphabet. Romanization has been around for a long time. For Japanese, the Hepburn Romanization system was first presented in 1887. The Hepburn Romanization for the Japanese 'computer' above is 'kompyuta'. The Hepburn system is widely enough known that a PERL module for Hepburn is available from the CPAN archive.

In addition to Hepburn, there has been a long practice by the USA Library of Congress to Romanize foreign scripts when cataloging the titles of books written in foreign languages. Figure 1 presents a list of about 55 languages for which the Library of Congress has published Romanization tables. Note that major Indian subcontinent languages of Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu and Urdu are included. For example, the Cyrillic Клинтон or the Greek Κλίντον can easily be Romanized to Klinton. For Russian and Greek, the transformation is usually reversible. For the major Indian language, Hindi, it is easily possible to find the translation for Clinton, but for the south Indian language of Tamil, translations are less easily found. Yet Tamil is a rather regular phonetic language and

foreign names are often transliterated when news stories are written in Tamil. Figure 2 is a translated news story in Tamil, when the main names (Presidents Clinton and Yeltsin) are Romanized.

2. TRANSLITERATION/ROMANIZATION

In the sweep of methods for recognition of out-of-vocabulary terms between languages and for automatic phonetic recognition of borrowed terms, Romanization has become a much-neglected stepchild. However phonetic transliteration (and back-transliteration from the target language to the source language) requires large training sets for machine learning to take place. For less-commonly taught languages, such as, for example, Indian subcontinent languages, such training sets may not be available. Romanization, on the other hand, requires that rules for alphabet mapping be already in place, developed by experts in both target and source languages. However, once the target language word has been rendered into its Latin alphabet equivalent, we still have the problem of matching it to its translation in the source.

3. APPROXIMATE STRING MATCHING

Once one has Romanized a section of non-English text containing OOV, the task remains to find its English word equivalents. The natural way to do this is using approximate string matching techniques. The most well-known technique is edit distance, the number of insertions, deletions and interchanges necessary to transform one string to its matching string. For example, the edit distance between computer and kompyuta (コンピュータ) is 5. Easier to comprehend is between English and German, where the Edit distance between fish (E) and fisch (DE) is 1. However, the edit distance between fish(E) and frisch (DE) is 2, whereas between the correct translations fresh (E) and frisch (DE) is also 2. Thus Martin Braschler of the University of Zurich has remarked, "Edit distance is a terrible cross-lingual matching method." Approximate string matching has a lengthy history for both fast file search techniques as well as finding matches of minor word translation variants across languages. Q-grams, as proposaed by Ukkonen (1992) counts the number of substrings of size 'q' in common between the strings being matched. A variant of q-grams are targeted s-grams where q is of size 2 and *skips* are allowed to omit letters from the match. Pirkkola and others (2003) used this technique for cross-language search between Finnish, Swedish and German. Using s-gram skips solves the fish – fisch differential above.

An alternative approach, which has been around for some time, is the Phonix method of Gadd (1998) which applies a series of transformations to letters (for example, c → k, in many cases, e.g. Clinton → Klinton) and shrinks out the vowels, (Clinton → Klntn). If we apply this transformation to the English Japanese above, we have computer → kmptr and compyuta →kmpt. The original version of Phonix only kept the leading four resulting characters, and would result in an exact match. Zobel and Dart (1995) did an extensive examination of approximate matching methods for digital libraries and their second paper (1996) proposed an improved Phonix method they titled Phonix-plus which did not truncate to 4 characters, but instead rewarded matches at the beginning. They combined this with edit distance for the Zobel-Dart matching algorithm.

4. SUMMARY AND POSITION

The current fashion for utilizing statistical machine learning as the solution to all problems in machine translation has led to the neglect of rule-based methods which, this paper argues, are both well-developed and could complement statistical approaches. Romanization would work especially well for non-Latin scripted languages for which training corpora are limited. The approach has two steps: 1) Romanization of the script using well-documented methods, followed by 2) Approximate string matching between Romanized words in the target language and possible translation candidates in the source language.

5. ACKNOWLEDGMENTS

Much of this work was originally done while the author was a visiting researcher at the National Institute of Infomatics (NII) in Tokyo in the summer of 2007 supported by a grant from NII.

language. So we ask: Is there a place for Romanization in CLIR? And how can it be exploited? The key is the examination of approximate string matching methods to find the correspondences between words of the target and source languages.

6. REFERENCES

[1] Knight, K and J Graehl (1997), Machine Transliteration, Association for Computational Linguistics (1997): ???-???. http://www.ala.org/ala/acrl/acrlpubs/crljournal/collegeresearch.cfm

[2] T. Gadd (1988), Fisching fore Werds: Phonetic Retrieval of Written Text in Information Systems***, Program, 22(3):222–237, 1988***

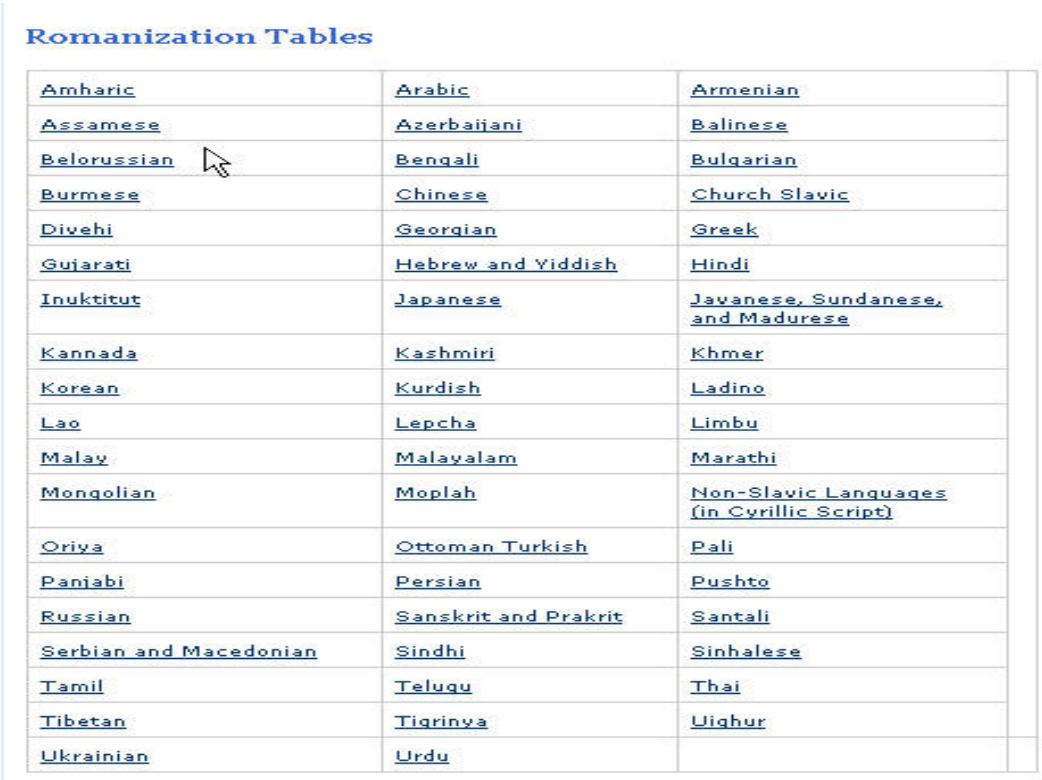
[3] R. Steinberger and B. Pouliquen (2007). Cross-lingual named entity recognition. Special issue of Lingvistic Investigationes, 30:135–162, 2007.

[4] J. Zobel and P. Dart (1995). Finding approximate matches in large lexicons. Softw. Pract. Exper., 25(3):331–345, 1995.

[5] J. Zobel and P. Dart (1996). Phonetic string matching: lessons from information retrieval. In SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 166–172, New York, NY, USA, 1996. ACM Press.

[6] A. Pirkola, J. Toivonen, H. Keskustalo, K. Visala, and K. Jarvelin (2003). Fuzzy translation of cross-lingual spelling variants. In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 345–352, New York, NY, USA, 2003. ACM Press.

[7] E. Ukkonen (1992). Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science 92 (1992), 191-21



Romanization Tables

Amharic	Arabic	Armenian
Assamese	Azerbaijani	Balinese
Belorussian	Bengali	Bulgarian
Burmese	Chinese	Church Slavic
Divehi	Georgian	Greek
Gujarati	Hebrew and Yiddish	Hindi
Inuktitut	Japanese	Javanese, Sundanese, and Madurese
Kannada	Kashmiri	Khmer
Korean	Kurdish	Ladino
Lao	Lepcha	Limbu
Malay	Malayalam	Marathi
Mongolian	Moplah	Non-Slavic Languages (in Cyrillic Script)
Oriya	Ottoman Turkish	Pali
Panjabi	Persian	Pushto
Russian	Sanskrit and Prakrit	Santali
Serbian and Macedonian	Sindhi	Sinhalese
Tamil	Telugu	Thai
Tibetan	Tigrinya	Uighur
Ukrainian	Urdu	

Figure 1: Library of Congress Romanization Language List

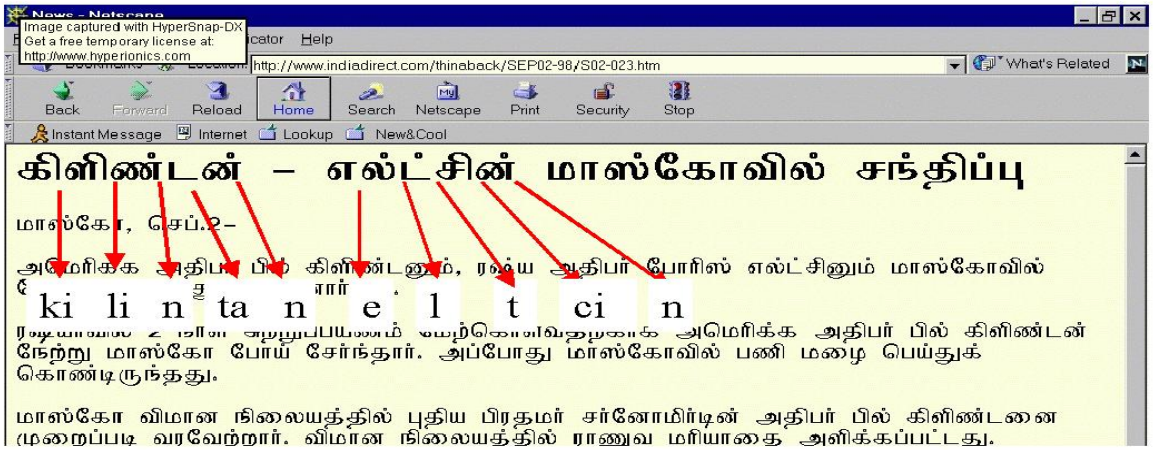


Figure 2: News story in the Tamil language of Clinton-Yeltsin Meeting, showing name Romanization

What's wrong with Cross-Lingual IR ?

John I. Tait
Information Retrieval Facility
Eschenbachgasse 11 Stg. 3
1010 Vienna, Austria
+43 1 236 94 74 6053
john.tait@ir-facility.org

ABSTRACT
To date much cross-language information retrieval research has focused on evaluation paradigms which were developed for monolingual web search. The paper argues that rather different scenarios are required for situations where cross-lingual search is a real requirement. In particular cross-lingual search is usually a collaborative as opposed to individual activity, and this needs to be taken into account in the evaluation of cross-lingual retrieval, especially when considering the notion of relevance.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]

General Terms
Documentation, Experimentation, Human Factors,

Keywords
Patent Search; Intellectual property Search; Information Retrieval; Cross-lingual Retrieval.

1. INTRODUCTION
It seems to me that for non-professional searchers there is very little requirement for cross lingual searching. Most non-professional searchers formulate queries in their native language and require results in that language. Even with much better machine translation than has ever been available before one would rarely find an automatic translation that one could include in ones school homework!

On the other hand professional searchers in field like Intellectual Property, Competitor Analysis, opinion mining and some international aspects of legal search, for example really do need Cross Lingual Information Retrieval (CLIR).

This paper outlines one such setting (patent search) and points out some problems with evaluation in that setting (especially the need for a sophisticated notion of relevance.

2. RELEVANCE IN CLIR
Experience with patent search has made it clear that while professional patent searchers need to access information in all languages in which patents can be filed: they require output in comparatively few languages: possibly only English and Chinese. This has implications for the design of cross lingual information systems, but also the evaluation including the ways in which relevance is judged.

This brief paper is not the place to present a detailed description of professional patent search in practice but see Hunt, Nguyen and Rogers [1] for example for more information including taxonomies of patent search.

Generally patent searchers will be instructed by a patent attorney acting on behalf an inventor or their employer. More complex searches might be done at the behest of strategic business managers requesting patent landscape searching to determine, for example whether research and development investment in a particular area is likely to yield patentable results.

The patent searcher will then formulate a series of queries (the search strategy) which will be addressed to often several different search systems. In practice most searching is on English abstracts, but really thorough searching for patentability for example requires searching of many different languages. This is a high recall task, in which it is important not to miss relevant documents.

Now there are several steps in the judgement of relevance in this context. First, the searcher needs make initial judgements of the relevance of patents (and indeed scientific articles and other material which may show the patent is not original, or obvious for instance). Then the patent attorney will review the results of the search; and in some cases other people: for example technical specialists (chemical engineers, molecular geneticists, search engine engineers etc.), language specialist, other lawyers, business managers and so on.

Now each of these groups, and the group collectively for an individual search, will bring different judgements of relevance to the retrieved document set. This needs to be taken into account and modelled explicitly in the evaluation.

Consider potential confounding factors in the experiment: what we are attempting to judge is the ability of the searcher to use the system to locate and determine the relevance (as assessed by the whole group). Quality of result translation may, for example, cause incorrect determination of relevance (or irrelevance) and we really need evaluation frameworks which take this into account.

Now I'm not claiming to say much new here: See Saracevic [2] for much more sophisticated approach: but those ideas do need to be more rigorously and consistently applied to CLIR evaluation.

3. OTHER ASPECTS OF EVALUATION
The consideration of confounding factors in our evaluation experiments leads onto some more general requirements of evaluations of CLIR for professionals search. It is not appropriate to give an exhaustive list here, but factors to be taken into account include:

1. The place of the computer systems in the human system;
2. The need for component based evaluation;
3. The need to assess the impact of frozen collections on the ecological validity of the experiment.

All this needs more careful thinking through than has been done to date.

4. CONCLUSION

Conventional CLIR evaluations have relied very much on the Cranfield experimental model pioneered by Cyril Claverdon, Karn Sparck Jones and others [3]. This paper is really a plea to move to more sophisticated models of evaluation fo professional serach, the context in which cross lingual retrieval is really valuable.

5. ACKNOWLEDGMENTS

I would like to thank my colleagues on the evaluation working group at the recent Interactive Information Retrieval Dagstuhl who developed my thinking on this topic; my colleagues in Matrixware and the IRF especially those who have worked on the CLEF IP and TREC CHEM tracks; the many IP professionals who have taken the time to educate me about patent search – especially Henk Tomas.

6. REFERENCES

[1] Hunt, D. Nguyen, L., and Rodgers, M. Patent Search: Tools and Techniques. Wiley, 2007.

[2] Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. Journal of the American Society for Information Science and Technology, 58(3), 1915-1933.

[3] Jones Sparck, K. 1981 *Information Retrieval Experiment*. Butterworth-Heinemann.

User Study of the Assignment of Objective and Subjective Type Tags to Images in Internet, considering Native and non Native English Language Taggers

David Nettleton
Pompeu Fabra University
Tanger, 122-140
08018 Barcelona, Spain
+34 93 542 25 00
david.nettleton@upf.edu

Mari-Carmen Marcos
Pompeu Fabra University
Roc Boronat,138
08018 Barcelona, Spain
+34 93 542 13 10
mcarmen.marcos@upf.edu

Bartolomé Mesa
Autonomous University of Barcelona
Edifici K – Campus UAB
08193 Barcelona, Spain
+34 93 581 1876
barto.mesa@uab.cat

ABSTRACT
Image tagging in Internet is becoming a crucial aspect in the search activity of many users all over the world, as online content evolves from being mainly text based, to being multi-media based (text, images, sound, …). In this paper we present a study carried out for native and non native English language taggers, with the objective of providing user support depending on the detected language skills and characteristics of the user. In order to do this, we analyze the differences between how users tag objectively (using what we call 'see' type tags) and subjectively (by what we call 'evoke' type tags). We study the data using bivariate correlation, visual inspection and rule induction. We find that the objective/subjective factors are discriminative for native/non native users and can be used to create a data model. This information can be utilized to help and support the user during the tagging process.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Indexing methods.

General Terms
Measurement, Experimentation, Human Factors.

Keywords
Image tagging, tag recommendation, user support, statistical analysis, user study.

1. INTRODUCTION
The English language is widely used in Internet, although for many of the people who use English in Internet, it is not their native language. In the image tagging context, when a non-native English tagger defines tags for an image, due to their limited knowledge of the language they may define incorrect tags or tags for which there exists a better word. In this paper, we will consider some of the difficulties for non-native English taggers and how to offer them appropriate help, such as tag word recommendation.

In order to do this, we derive factors to identify differences between how users tag objectively (using what we call 'see' type tags) and subjectively (by what we call 'evoke' type tags). The hypothesis is that 'evoke' (subjective) tags require more skill and knowledge of vocabulary than 'see' (objective) tags. Therefore, the tagger, and especially the non-native tagger, will require additional help for this type of tags.

We have collected information in a custom made website and questionnaire, from tag volunteers in two different countries (Spain and the United States), for native/non native speakers in the English language.

2. STATE OF THE ART AND RELATED WORK
We ask up to what point users with different language skill levels vary in their way of indexing contents which are similar or the same. Specifically, we will look at the description of images, and the difference between tags (labels) which represent feelings, emotions or sensations compared with tags which represent objective descriptions of the images [2][5]. As a point of reference, we consider the popular Flickr (http://www.flickr.com) website. The images published in Flickr can be labeled or tagged (described using labels or tags) by the same author and also by the rest of the users of this service.

In recent years tag recommendation has become a popular area of applied research, impulsed by the interests of major search engine and content providers (Yahoo, Google, Microsoft, AOL, ..). Different approaches have been made to tag recommendation, such as that based on collective knowledge [8], approaches based on analysis of the images themselves (when the tags refer to images) [1], collaborative approaches [6], a classic IR approach by analyzing folksonomies [7], and systems based on personalization [3]. With respect to considerations of non-native users, we can cite works such as [10]. In the context of tags for

blogs, [6] used filter detection to only choose English language documents/tags. Finally we can cite approaches based on complex statistical models, such as [9].

In conclusion of the state of the art, to the best of our knowledge there are non or few investigators working on support for non-native taggers of images, and making the distinction and support for subjective versus objective tagging, which are two of the main lines of our work presented in this paper.

3. METHODOLOGY – DESIGN OF EXPERIMENTS FOR USER EVALUATION

For this study we have selected 10 photographs from Flickr. The objective of each image is to evoke some type of sensation. The 10 photographs we have used have been chosen for their contrasting images and for their potential to require different tags for 'see' and 'evoke'. Image 1 is of a person with his hands to his face; Image 2 is of a man and a woman caressing; Image 3 is of a small spider in the middle of a web; Image 4 is of a group of natives dancing in a circle with a sunset in the background; Image 5 is of a lady holding a baby in her arms; Image 6 is of a boy holding a gun ; Image 7 is of an old tree in the desert, bent over by the wind; Image 8 is of a hand holding a knife; Image 9 is a photo taken from above of a large cage with a person lying on its floor; finally, Image 10 is of a small bench on a horizon.

We have created a web site with a questionnaire in which the user introduces his/her demographic data, their tags for the photographs (tag session) and some questions which the user answers after completing the session. The capture of tag sessions has been carried out for native and non-native English, and our website reference is:

http://www.tradumatica.net/bmesa/interact2007/index_en.htm .

Tag Session Capture. During a tag session the users must assign between 4 and 10 tags which are related to the objects which they can see in the image and a similar number of tags related to what each image evokes for them, in terms of sensations or emotions. With reference to Figure 1, in the first column the user writes the tags which express what they see in the image, while in the second column the user writes the tags which describe what the image evokes. We have currently accumulated a total of 162 user tag sessions from 2 different countries, involving the tasks of description of the photographs in English. For approximately half of the users, English is their native language and for the other half it is a second language.

Data and Factors for Analysis. From the tags collected and the information which the users have provided, we can compare results in the English language used by native and non natives in that language. Our data is captured from taggers in the United States (native) and from Spain (non native). For each tag session, we collect the following information: language in which the tag session is conducted; easiest image to tag (user is asked); most difficult image to tag (user is asked); the tags themselves assigned for each image, for "See" and "Evoke" separately, and the order in which the tag is assigned. We also record the type of language (if the current tagging language is native or not for the user).

Figure 1. Example of how the user enters the tags for a given image.

The following factors were derived from the tagging session data (statistically averaged and grouped by user and image):

- Easiness: average number of tags used for "see" and "evoke". This value is compared with the question which refers to the ease or difficulty which a user had to tag the image for "see" and in "evoke". One assumption is that the images evaluated as easier to tag should have more tags. Also, users who possess a greater descriptive vocabulary in the tagging language should define a greater number of tags.

- Similarity: frequency of the tags used for "see" and for "evoke". The tags which present a greater frequency in each image will be compared to detect similarities or differences between native and non-native taggers.

- Spontaneity: tags used as first option for "see" and for "evoke". The tags which appear as first option in each image will be compared to detect similarities or differences between native and non-native taggers.

4. DATA PROCESSING

The following factors were derived from the tag session data:

"Easiness" is represented by the following six factors: "anumTagsSee", "anumTagsEvoke", "asnumTermsSee", "asnumTermsEvoke", "aanumTermsSee" and "aanumTermsEvoke". These factors represent, respectively, the average number (for all images) of tags used for "See", the average number (for all images) of tags used for "Evoke", the average of the sum (for each image) of the number of terms used in each tag for "See", the average of the sum (for each image) of the number of terms used in each tag for "Evoke", the average number of terms (for each tag) used for "See" tags and the average number of terms (for each tag) used for "Evoke" tags. We recall

that all these values are summarized by image and user, and that a tag consists of one or more terms (individual words).

"**Similarity"** is represented by the following four factors: "asimSee", "asimEvoke", "atotSimSee" and "atotSimEvoke". The factor "aSimSee" represents the average similarity of a given tagging of an image by a given user for "See", in comparison with all other taggings of the same image by all other users. This is essentially a frequency count of tag coincidences. The factor "aSimEvoke" represents the same statistic as "aSimSee", but calculated for the "Evoke" type tags. The factor "atotSimSee" is equal to "asimSee' divided by the number of users, which gives a sort of 'normalized' value. The factor "atotSimEvoke" represents the same statistic as "atotSimSee", but calculated for the "Evoke" type tags.

"Spontaneity" is represented by the following two factors: "aespSee" and "aespEvoke". The factor "aespSee" represents the spontaneity of a given tagging of an image in a given tag session for "See", by comparing it with the most frequent tags chosen as first option for the same Image. The factor "aespEvoke" represents the same statistic as "aespSee", but calculated for the "Evoke" type tags.

5. QUANTITATIVE EVALUATION

In this section we show results of the data analysis and data modeling using the IM4Data (IBM Intelligent Miner for Data V6.1.1) Data Mining tool [4].

Data Analysis – Statistical Methods and Visualization. Figures 2 and 3 are produced from the 'SessionD' dataset for native English taggers and non-native taggers, respectively. They are ordered by the Chi-squared statistic relative to the 'typeLanguage' label. We recall that this dataset contains attributes which represent the 'easiness', 'similarity' and 'spontaneity' factors for the user tag sessions. Refer to the definitions of these factors in Sections 3 and 4 of the paper. We observe that the first four ranked attributes in Figure 2 (native) and Figure 3 (non native) are 'atotSimEvoke', 'mostDifficult', 'asimEvoke' and 'aespSee', although the ordering is different for attributes 2 to 4. From this we observe that two of the attributes most related to the native/non native label (as indicated by Chi-Squared) are variables related to the similarity of the evoke type tags. This is coherent with the hypothesis that non native users will find it more difficult to think of vocabulary to define emotions. If we look at the distributions of 'atotsimEvoke' and 'asimEvoke' in Figures 2 and 3, we see that the non-natives (Figure 3) have a greater frequency in the higher (rightmost) part of the distribution, which means that there is more coincidence between the non-native tags, and therefore less diversity.

Rule Extraction. The IM4Data tree/rule induction algorithm was used for data modeling. For testing, we have manually created test datasets using a 5x2-fold cross-validation. We used 14 input attributes: easiest, mostDifficult, anumTagsSee, anumTagsEvoke, asnumTermsSee, asnumTermsEvoke, aanumTermsSee, aanumTermsEvoke, asimSee, asimEvoke, atotSimSee, atotSimEvoke, aespSee, aespEvoke; and one output attribute (class): 'typeLanguage'.

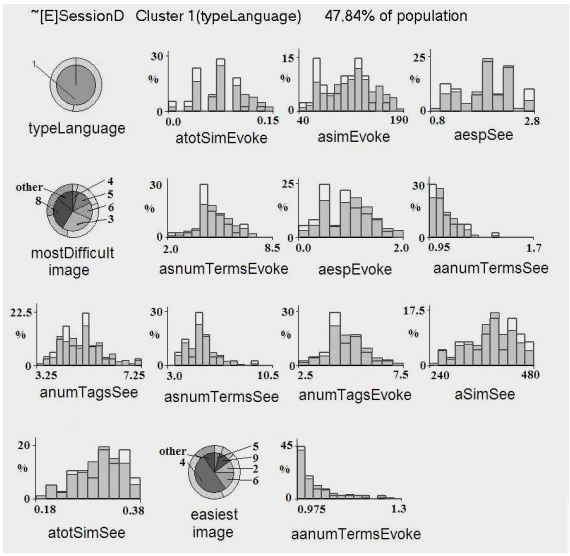


Figure 2. Distributions of variables of dataset 'SessionD', for native English taggers.

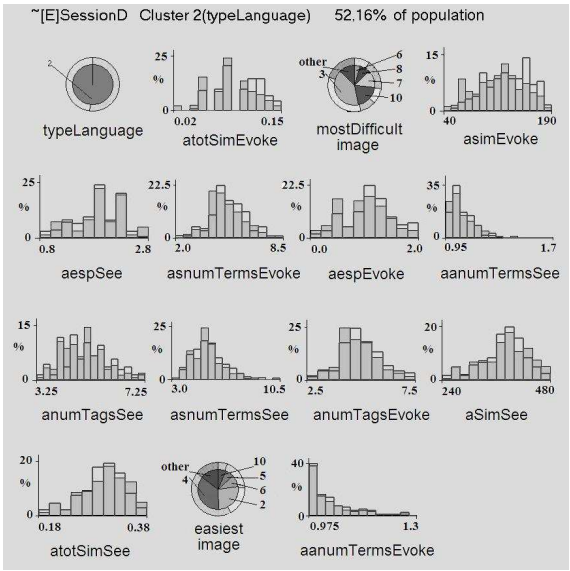


Figure 3. Distributions of variables of dataset 'SessionD', for non-native taggers.

With reference to Figure 4, we see the pruned tree induced by IM4Data on the SessionD dataset, including the details of the decision nodes and classification nodes. We observe that attributes 'asimEvoke' and 'mostDifficult' have been used in the upper part of the tree (asimEvoke < 138.15, mostDifficult in [image9, image3, ,image10, image7]). Thus, they represent the most general and discriminatory factors to classify 'typeLanguage', that is the native and non-native users. We note that lower down in the tree the attribute 'asnumTermsSee' has been used.

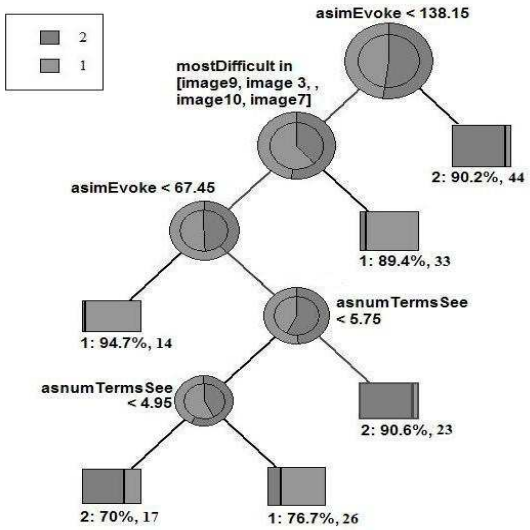


Figure 4. Pruned Classification Tree: dataset 'SessionD'.

Table 1. 'SessionD': test precision for 5x2 fold cross validation

	native†	non-native††	MP*
fold1	65.5, 21.1	78.9, 34.5	71.08
fold2	88.3, 32.2	67.8, 11.7	77.07
fold3	85.2, 33.9	66.1, 14.3	76.17
fold4	70.6, 34.4	65.6, 29.4	77.60
fold5	89.6, 35.0	65.0, 10.4	76.42
Geometric mean for folds	79.2, 30.8	68.5, 17.7	75.63

*MP=Model Precision †{%Rate: True Positive, False Positive}, ††,{%Rate: True Negative, False Negative}

With reference to Table 1, we present the test results (test folds) for the tree induction model built from the SessionD factors. The overall precision of the model over 5 folds is 75.63%. The low percentage of false positives and false negatives over the five folds indicates that we have a 'robust' model. We conclude from the results that with the derived factors for 'Easiness', 'Similarity' and 'Spontaneity' we are able to produce an acceptably precise model (75.63%), using real data and 'typeLanguage' as the output class. This model distinguishes between English native and non-native taggers, based on the given input variables and derived factors.

6. TAG RECOMMENDATION

Recommendation of 'evoke' tags based on 'see' tags: if the user has already defined the 'see' tags, then the system can recommend the 'evoke' tags, based on the 'see' tags. For example, with reference to the list of most frequent 'see' and 'evoke' tags for Image 10 (Section 3), if the non native user defines the following 'see' tags: 'sky', 'grass' and 'bench', then the system would consult a dictionary of 'see' tags and corresponding 'evoke' tags which have been defined previously by other (native or more highly skilled) users.

7. CONCLUSIONS

As a conclusion from the present work and the available data and derived factors, we can reasonably infer that there is a significant difference between "see" and "evoke" type tags, and we have successfully built a data model from these factors (Figure 4, Table 1). We have determined that native and non native taggers have distinctive characteristics in terms of the tag type based on objective or subjective tagging. Some interesting results were also found with respect to the easiest and most difficult images, differentiating between native and non native taggers.

8. REFERENCES

[1] Anderson, A., Raghunathan. K., Vogel, A., 2008. TagEz: Flickr Tag Recommendation. Association for the Advancement of Artificial Intelligence (www.aaai.org). http://cs.stanford.edu/people/acvogel/tagez/

[2] Boehner, K., DePaula R., Dourish, P., Sengers, P., 2007. How emotion is made and measured. Int. Journal of Human-Computer Studies. 65:4, 275-291.

[3] Garg, N., Weber, I., 2008. Personalized, interactive tag recommendation for flickr. Proceedings of the 2008 ACM conference on Recommender Systems, Lausanne, Switzerland. pp. 67-74, ISBN:978-1-60558-093-7.

[4] Im4Data, 2002. Using the Intelligent Miner for Data V8 Rel. 1. IBM Redbooks, SH12-6394-00.

[5] Isbister, K., Hook, K., 2007. Evaluating affective interactions. Int. Journal of Human-Computer Studies, 65:4, 273-274.

[6] Lee, S. A, 2007. Web 2.0 Tag Recommendation Algorithm Using Hybrid ANN Semantic Structures. Int. Journal of Computers, Issue 1, Vol. 1, 2007, pp. 49-58. ISSN: 1998-4308.

[7] Lipczak, M., Angelova, R., Milios, E., 2008. Tag Recommendation for Folksonomies Oriented towards Individual Users. ECML PKDD Discovery Challenge 2008, Proc. of WWW 2008.

[8] Sigurbjörnsson, B., van Zwol, R., 2008. Flickr Tag Recommendation based on Collective Knowledge. WWW 2008, Beijing, China, ACM 978-1-60558-085-2/08/04.

[9] Song, Y. , 2008. Real-time Automatic Tag Recommendation. SIGIR'08, July 20–24, 2008, Singapore. ACM 978-1-60558-164-4

[10] Sood, S.C., Hammond, K., Owsley, S.H., Birnbaum, L., 2007. TagAssist: Automatic Tag Suggestion for Blog Posts. Int. Conf. on Weblogs and Social Media (ICWSM), 2007, Boulder, Colorado, USA.

Multilingual Wikipedia, Summarization, and Information Trustworthiness

Elena Filatova
Fordham University
Department of Computer and Information Sciences
filatova@cis.fordham.edu

ABSTRACT

Wikipedia is used as a corpus for a variety of text processing applications. It is especially popular for information selection tasks, such as summarization feature identification, answer generation/verification, etc. Many Wikipedia entries (about people, events, locations, etc.) have descriptions in several languages. Often Wikipedia entry descriptions created in different languages exhibit differences in length and content. In this paper we show that the pattern of information overlap across the descriptions written in different languages for the same Wikipedia entry fits well the *pyramid* summary framework, i.e., some information facts are covered in the Wikipedia entry descriptions in many languages, while others are covered in a handful number of descriptions. This phenomenon leads to a natural summarization algorithm which we present in this paper. According to our evaluation, the generated summaries have a high level of user satisfaction. Moreover, the discovered pyramid structure of Wikipedia entry descriptions can be used for Wikipedia information trustworthiness verification.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Measurement, Experimentation, Human Factors

Keywords

Wikipedia, summarization, multilinguality

1. INTRODUCTION

"Wikipedia is a free, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation."[1,2] It provides descriptions of people, events, locations, etc. in many languages. Despite the recent discussion of the Wikipedia descriptions trustworthiness or lack of thereof [9], Wikipedia is widely used in information retrieval (IR) and natural language processing (NLP) research. Thus, the question arises what can be done to increase the trustworthiness to the information extracted from Wikipedia. We believe, Wikipedia itself has resources to increase its trustworthiness.

Most Wikipedia entries have descriptions in different languages. These descriptions are *not* translations of a Wikipedia entry description from one language into other languages. Rather, Wikipedia entry descriptions in different languages are independently created by different users. Thus, the length of the entry descriptions about the same Wikipedia entry varies greatly from language to language. Obviously, texts of different length cannot contain the same amount of information about an entry.

In this paper we compare descriptions of Wikipedia entries written in different languages and investigate the pattern of information overlap. We show that information overlap in entry descriptions written in different languages corresponds well to the pyramid summarization model [15, 11]. This result helps the understanding of the combined value of the multilingual Wikipedia entry descriptions. On the one hand, multilingual Wikipedia provides a natural summarization mechanism. On the other hand, to get a complete picture about a Wikipedia entry, descriptions in all languages should be combined. Finally, this pyramid structure can be used for information trustworthiness verification.

The rest of the paper is structured as follows. In Section 2 we describe related work including work on utilizing Wikipedia and on analyzing Wikipedia information trustworthiness. In Section 3 we provide a motivation example for our research. In Section 4 we describe our corpus, the summarization-based experiments we ran to analyze multilingual Wikipedia information overlap; and discuss the results of these experiments. In Section 5 we draw conclusions from these experiments. In Section 6 we outline the avenues for future research.

2. RELATED WORK

Multilingual aspect of Wikipedia is used for a variety of text processing tasks. Adafre *et al.* [8] analyze the possibility of constructing an English-Dutch parallel corpus by suggesting two ways of looking for similar sentences in Wikipedia pages (using matching translations and hyperlinks). Richman *et al.* [12] utilize multilingual characteristics of Wikipedia to annotate a large corpus of text with Named Entity tags. Multilingual Wikipedia is used to facilitate cross-

[1] `http://en.wikipedia.org/wiki/Wikipedia`

[2] Wikipedia is changing constantly. All the quotes and examples from Wikipedia presented and analyzed in this paper were collected on February 10, 2009, between 14:00 and 21:00 PST.

SIGIR Workshop on Information Access in a Multilingual World '09 Boston, Massachusetts USA

language IR [13] and to perform cross-lingual QA [6].

The described applications do not raise a question on whether the information presented in Wikipedia articles is trustworthy. Currently, the approaches to rate the trustworthiness of Wikipedia information are dealing with the text written in only one language.

Wikipedia content trustworthiness can be estimated using a combination of the amount of the content revision and the author reputation performing this revision [2]. Wikipedia author reputation in its turn can be computed according to the content amount that is preserved for a particular author by other authors [3]. Another way to use edit history to estimate information trustworthiness is to treat Wikipedia article editing as a dynamic process and to use dynamic Bayesian network trust model that utilized rich revision information in Wikipedia for trustworthiness estimation [16].

Another approach suggested to estimate Wikipedia trustworthiness is to introduce an additional tab to the Wikipedia interface *Trust tab*. This tool enables users to develop their own opinion concerning how much and under what circumstances, they should trust entry description information [10].

The research closest to ours was recently described in Adar *et al.* [1] where the main goal is to use self-supervised learning to align or/and create new Wikipedia infoboxes across four languages (English, Spanish, French, German). Wikipedia infoboxes contain a small number of facts about Wikipedia entries in a semi-structured format. In our work, we deal with plain text and disregard any structured data such a infoboxes, tables, etc. It must be noted, that the conclusions that are reached in parallel for structured Wikipedia information by Adar *et al.* and for unstructured Wikipedia information by us are very similar. These conclusions stress the fact that the most trusted information is repeated in the Wikipedia entry descriptions in different languages. At the same time, no single entry descriptions can be considered as the complete source of information about a Wikipedia entry.

3. INFORMATION OVERLAP

Currently, Wikipedia has entry descriptions in more than 200 languages. The language with the largest number of entry descriptions is English [8, 5] but the size of non-English Wikipedia is growing fast and represents a rich corpus.[3]

Most existing NLP applications that use Wikipedia as the training corpus or information source assume that Wikipedia entry descriptions in all languages are a reliable source of information. However, according to our observations, Wikipedia descriptions about the same entry (person, location, event, etc.) in different languages frequently cover different sets of facts. Studying these differences can boost the development of various NLP applications (i.e., summarization, QA, new information detection, machine translation, etc.). According to the Wikipedia analysis [7], there are two major sources of differences in the descriptions of the same Wikipedia entry written in different languages:

- the amount of information covered by a Wikipedia entry description;[4]
- the choice of information covered by a Wikipedia entry description.

In this paper we analyze the information overlap in Wikipedia entry descriptions written in several languages.

For example, baseball is popular in the USA, Latin America, and Japan but it is not in Europe or Africa. Wikipedia has descriptions of *Babe Ruth* in 18 languages: the longest and most detailed descriptions are in English, Spanish and Japanese. The description of *Babe Ruth* in Finnish has five and in Swedish - four sentences. These short entry descriptions list several general biographical facts: dates of birth, death; the fact that he was a baseball player. It is likely, that the facts from the Swedish and Finnish entry descriptions about *Babe Ruth* will be listed in a summary of the English language Wikipedia entry description of him.

4. CORPUS ANALYSIS EXPERIMENT

In this paper, we investigate how the information overlap in multilingual Wikipedia can be used to create summaries of entry descriptions. Our results show that the information that is covered in more than one language corresponds well to the pyramid summarization model [15, 11].

4.1 Data Set

For our experiments, we used the list of people created for the Task 5 of DUC 2004: biography generation task (48 people).[5] We downloaded from Wikipedia all the entry descriptions in all the languages corresponding to each person from the DUC 2004 list. For our experiments we used Wikitext, the text that is used by Wikipedia authors and editors. Wikitext can be obtained through Wikipedia dumps.[6] For our experiments we removed from the wikitext all the markup tags and tabular information (e.g., infoboxes and tables) and kept only plain text. There is no commonly accepted standard wikitext language, thus our final text had a certain amount of noise which, however, as discussed in Section 5, did not affect our experimental results.

For this work, for each Wikipedia entry (i.e., DUC 2004 person) we downloaded corresponding entry descriptions in all the languages, including Esperanto, Latin, etc. To facilitate the comparison of entry descriptions written in different languages we used the Google machine translation tool[7] to translate the downloaded entry descriptions into English. The number of languages covered currently by the Google translation system (41) is less than the number of languages used in Wikipedia (265). However, the language distribution in the collected corpus corresponds well the language distribution in Wikipedia and the collected Wikipedia subset can be considered a representative sample [7].

Five people from the DUC 2004 set had only English Wikipedia entry descriptions: *Paul Coverdell*, *Susan McDougal*, *Henry Lyons*, *Jerri Nielsen*, *Willie Brown*. Thus, they were excluded from the analysis. The person whose Wikipedia entry had descriptions in most languages (86) was *Kofi Annan*. On average, a Wikipedia entry for a DUC 2004 person had descriptions in 25.35 languages. The description in English was not always the longest description: in 17 cases the longest description of a Wikipedia entry for a DUC 2004 person was in a language other than English.

4.2 Data Processing Tools

After the Wikipedia entry descriptions for all the DUC 2004 people were collected and translated, we divided these descriptions into sentences using the LingPipe sentence chun-

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] In this work, the length of a Wikipedia entry description is measured in sentences used in the text description of a Wikipedia entry.

[5] http://duc.nist.gov/duc2004/tasks.html/

[6] http://download.wikimedia.org/

[7] http://translate.google.com/

Algorithm	
1	Submit the person's name to Wikipedia
2	Get Wikipedia entry descriptions for this person in all possible languages
3	Remove non-plain text information from the descriptions
4	For all the languages handled by the Google MT, translate entry descriptions into English
5	Break English texts into sentences
6	Use a similarity measure to identify what English sentences have counterparts in entry descriptions in other languages
7	Rank all the sentence from the English document according to the number of languages that have similar sentences
8	If several sentences are placed on the same level, list these sentence in the order they appear in the Wikipedia entry description in English
9	Use the top three levels from the above ranking

Table 1: Algorithm outline.

ker [4]. For each DUC 2004 person we compared a description of this person in English against the descriptions of this person in other languages that were handled by the Google translation system. We counted descriptions in how many languages had sentences corresponding to the sentences in the description in English. To identify matching sentences we used the LingPipe string matching tool based on TF/IDF and cosine measure of angular similarity over dampened and discriminatively weighted term frequencies. We used three similarity thresholds: 0.5, 0.35, 0.2.

4.3 What was Measured

To evaluate how much information is repeated in the descriptions of the same person in different languages we measured similarity of the person's description in English and in other languages. For the 0.5 similarity threshold, the sentences marked as similar were almost identical. Using the 0.2 threshold allowed to search for non-identical sentences that still had a substantial word overlap.

Our hypothesis is that those facts (sentences) that are mentioned in the descriptions of a person in different languages fit well the pyramid summarization model. For example, if we are to summarize a description of a person from the English Wikipedia: first, we should add to the summary those sentences that have their counterparts in the most number of descriptions of this person in the languages other than English. Sentences added on this step correspond to the top level of the pyramid. If the length of the summary is not exhausted then, on the next step, we add to the summery those sentence that appear in the next most number of languages, and so on. Thus, we can place sentences on different levels of the pyramid, with the top level being populated by the sentences that appear in the most languages and the bottom level having sentences that appear in the least number of languages. For our experiments we used sentences from the top three levels of this pyramid. All the sentences added to the summary should appear in at least two languages other than English. Table 1 has a schematic outline of the described algorithm.

4.4 Example and Experiment Discussion

Table 2 presents three-level summaries for the English Wikipedia description of *Gene Autry*. Wikipedia has descriptions of *Gene Autry* in 11 languages: in English and in ten other languages each of which can be translated into English by the Google Translation system.

Using similarity of 0.5 we get one sentence from the English description that has counterparts in at least two other languages (here, in three other languages). This sentence's ID is 0: it is the entry description introductory sentence.

Using similarity 0.35 we get a summary consisting of six sentences. The sentence on the top level is the same sentence that was listed in the previous summary. However, having a more permitting similarity threshold, this sentence was mapped to similar sentences in 7 languages, rather than in 3. Next level consists of those sentences from the English description that were mapped to sentences in three other languages. Sentences on the third level were mapped to sentences from two other languages. It is interesting to notice that the sentences included in the summary are coming from different parts of the document that has 88 sentences.

The summary created using the 0.2 threshold contains the introductory sentence as well as sentences not included in the summaries for the 0.5 and 0.35 similarity threshold.

Despite the fact that for our experiment we chose the set of people used for the DUC 2004 biography generation task, we could not use the DUC 2004 model summaries for our evaluation. These models were created using the DUC 2004 corpus, while in our experiments we used a subset of multilingual Wikipedia. Moreover, Wikipedia entry descriptions about the DUC 2004 people had dramatic updates since 2004. For example, *Jörg Haider* died of injuries from a car crash on October 11, 2008 and this information is included into our three-level summaries.

Due to space constraints, in this paper we report only the results obtained using similarity threshold of 0.35. Also, in the experiment described in this paper we analyze only those sentences from the English text that appear in at least two other languages, with the exception for *Louis Freeh*, for whom only one language was handled by the Google Translation system. Thus, the summary for the English entry description about *Louis Freeh* has only one level which has all the sentences from the English entry description that have their counterparts in the only available translation. Thus, using the DUC 2004 set we created:

- **one**-level summaries for 5 people;
- **two**-level summaries for 3 people;
- **three**-level summaries for 35 people.

The length of the created summaries is measured in sentences. Table 3 presents information about the average and maximal length of summaries for all three levels combined and for each level separately. The summaries that do not have Level 2 and/or Level 3 are excluded from the corresponding average and maximum value computation. According to the presented data, on average, the output three-level summaries are rather short, however, some summaries can be quite long. We believe that such a difference between the average and the maximal length is due to:

1. the length variation of the English Wikipedia entry descriptions;
2. the number variation of descriptions (languages) for each a person and the lengths of these descriptions.

To evaluate the output three-level summaries we used Amazon Mechanical Turk as a source of human subjects who can reliably evaluate certain NLP tasks [14]. For each of the 43 outputs (for 43 people from the DUC 2004 set) we recruited five human annotators. The annotators were provided with: the name of the person; link to the Wikipedia entry description about this person in English; three-level

#	Lang.	Sent. ID	Text
Similarity 0.5			
1	3	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as The Singing Cowboy on the radio, in movies and on television.
Similarity 0.35			
1	7	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as "The Singing Cowboy" on the radio, in movies and on television.
2	3	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas.
		13	His first hit was in 1932 with "That Silver-Haired Daddy of Mine," a duet with fellow railroad man, Jimmy Long.
3	2	3	After leaving high school in 1925, Autry worked as a telegrapher for the St. Louis-San Francisco Railway.
		14	Autry also sang the classic Ray Whitley hit "Back in the Saddle Again," as well as many Christmas songs including "Santa Claus Is Coming to Town," his own composition "Here Comes Santa Claus," "Frosty the Snowman," and arguably his biggest hit "Rudolph the Red-Nosed Reindeer."
		72	Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.
Similarity 0.2			
1	7	0	Orvon Gene Autry (September 29, 1907 – October 2, 1998) was an American performer, who gained fame as "The Singing Cowboy" on the radio, in movies and on television.
2	6	1	Autry, the grandson of a Methodist preacher, was born near Tioga, Texas.
		73	His death on October 2, 1998 came nearly three months after the death of another celebrated cowboy of the silver screen, radio, and TV, Roy Rogers.
3	5	21	From 1940 to 1956, Autry had a huge hit with a weekly radio show on CBS, "Gene Autry's Melody Ranch." His horse, Champion, also had a radio-TV series "The Adventures of Champion."
		72	Gene Autry died of lymphoma at age 91 at his home in Studio City, California and is interred in the Forest Lawn, Hollywood Hills Cemetery in Los Angeles, California.

Table 2: Three-level summaries for Gene Autry (*#*: the summary level; *Lang.*: number of languages that contain a similar sentence; *Sent. ID*: the position of the sentence in the English description; *Text*: the sentence itself).

	Three-level summary	Level one	Level two	Level three
Avg	3.74	1.02	1.58	1.63
Max	9	2	6	7

Table 3: Summaries length: average and maximal.

summary of this Wikipedia entry description. We asked our human annotators to answer the following questions:

- Do you agree that the sentences listed on Level 1 are a good summary of the Wikipedia entry description about *Person* (assume, the number of sentences in the summary cannot exceed the number of sentences listed on Level 1)?
- Assume that the summary of the Wikipedia entry description about *Person* can have as many sentences as listed on Level 1 and Level 2 combined. Do you agree that the sentences listed on Level 1 and Level 2 are a good summary?
- Assume that the summary of the Wikipedia entry description about *Person* can have as many sentences as listed on Level 1, Level 2, and Level 3 combined. Do you agree that the sentences listed on Level 1, Level 2, and Level 3 are a good summary?

If the summary did not have Level 2 and/or Level 3 sentences, the annotator was asked to skip answering the corresponding questions.

5. RESULTS

Table 4 summarizes the results of the three-level summaries evaluation. The **Goodness** measure shows how many (out of five) annotators agreed that the summary for a particular level, given the length constraint, was good. The numbers in the table show the number of summaries that were considered good for each level according to a particular level of goodness. As it is mentioned in Section 4.4 not all summaries have Levels 2 and 3 filled in; the *Number of summaries* column in Table 4 has this information.

According to Table 4, no summary on Level 1 was uniformly considered bad. One summary was considered bad by four out of five annotators. This was the summary for *Paul Wellstone* with Level 1 consisting only of one sentence. We analyzed this sentence and discovered that it was incorrectly truncated due to our sentence chunker error.

> [Paul David Wellstone (July 21, 1944 - October 25, 2002) was a two-term U.S.] [Senator from the U.S. state of Minnesota and member of the Democratic-Farmer-Labor Party, which is affiliated with the national Democratic Party.]

This sentence was broken into two sentences (identified above by the square brackets), and only the first portion of the sentence was added to the Level 1 summary. Despite the fact that this portion contains important biographical information, it cannot be used as a stand-alone sentence. According to our analysis, three out of seven summaries that were judged as bad by three out of five annotators had exactly the same problem of incorrect sentence segmentation forcing only portions of sentences to be added to the summaries.

In addition to asking annotators to judge the quality of the created summaries we welcomed our annotators to leave comments about the summaries they read. Several annotators noticed text preprocessing errors (e.g., leftovers from the Wikitext XML tagging), however, it did not seem to affect their judgement of the summary quality: all the summaries containing such tags were marked as good. Another set of observations concerned the type of facts included in the summaries. For example, one annotator pointed out that the sentences from the summary about *Abdullah Öcalan*

	Goodness						Number of
Levels	**5**	**4**	**3**	**2**	**1**	**0**	summaries
1	28	4	3	7	1	0	43
1,2	12	3	12	4	5	2	38
1,2,3	5	6	14	8	1	1	35

Table 4: Evaluation results (using Mechanical Turk).

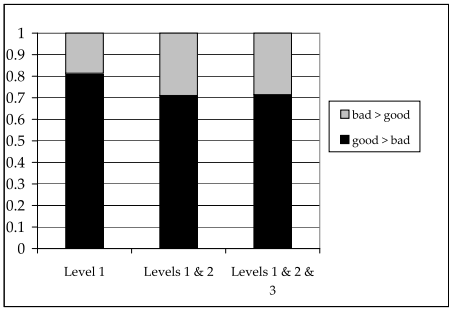


Figure 1: Combined results.

did not have **enough** information about his political activities and thus, the created summaries were judged as bad. Several annotators suggested that information about professional life of politicians would be more appropriate than the information about their private life (marriages). However, sentences containing information about private life were considered relevant and judged as good additions to summaries.

Figure 1 shows the combined numbers for Table 4. For each level we grouped all the numbers in two categories: those numbers where the majority of the annotators agreed that the summary was good and those numbers where the majority of the annotators decided that the summary was bad. As not all the summaries had sentences from all three levels, Figure 1 has encodes ratios rather than the absolute numbers listed in Table 4. This figure shows that overall the quality of the created summaries was quite high. In more than 80% of cases our annotators were happy with the summaries consisting of the Level 1 sentences, and in more that 70% of cases our annotators were happy with the summaries consisting of the sentences combined from Levels 1 and 2 and Levels 1, 2, and 3.

To conclude this section, we showed that information overlap in multilingual Wikipedia can be used for placing information facts into a pyramid structure. Thus, Wikipedia entry descriptions about the same entry in different languages can be treated as model descriptions with different foci depending on the personal, cultural, and other preferences of the Wikipedia contributors who create entry descriptions for different languages as well as on the community interest and attitude to the subject.The created Wikipedia entry pyramid can be also used to measure the entry information trustworthiness: the more descriptions mention a particular fact - the more trustworthy it is.

6. FUTURE WORK

We believe, studying Wikipedia multilinguality has a potential of providing training material for many NLP applications. For example, entry descriptions in different languages describing the same entry can be used as a training corpus for learning summarization features.

Also, we are interested in investigating how Wikipedia multilinguality can be used for opinion, contradiction and new information detection. As described in Section 3, the choice of information is critical and can be used as opinion marker.An important observation concerning the example presented in Section 3 is that irrespectively of the length of the descriptions of a person in different languages, none of these descriptions have any facts that contradict the facts in the descriptions of this Wikipedia entry in other languages. Rather, the discussed entry descriptions in different languages contain a subset of facts that appear in many languages plus, maybe, additional information. This allowed us to formulate and test the hypothesis that a set of Wikipedia entry descriptions about the same entry fits well the pyramid summarization model.

7. REFERENCES

[1] E. Adar, M. Skinner, and D. Weld. Information arbitrage in multi-lingual Wikipedia. In *WSDM*, 2009.

[2] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *WikiSym*, 2008.

[3] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW*, 2007.

[4] Alias-i. LingPipe 3.7.0. (accessed January 19, 2009), 2009. `http://alias-i.com/lingpipe`.

[5] L. Denoyer and P. Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.

[6] S. Ferrández, A. Toral, O. Ferrández, A. Ferrández, and R. Munoz. Applying Wikipedia's multilingual knowledge to cross-lingual Question Answering. *LNCS: Natural Language Processing and Information Systems*, 4592:352–363, 2007.

[7] E. Filatova. Directions for exploiting asymmetries in multilingual Wikipedia. In *CLIAW3*, 2009.

[8] S. Fissaha Adafre and M. de Rijke. Finding similar sentences across multiple languages in Wikipedia. In *Workshop on New Text – Wikis and blogs and other dynamic text sources*, 2006.

[9] A. Keen. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Bantam Books, 2007.

[10] D. McGuinness, H. Zeng, P. Pinheiro da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigations into trust for collaborative information repositories: A Wikipedia case study. In *MTW*, 2006.

[11] A. Nenkova, R. Passonneau, and K. McKeown. The Pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4(2), 2007.

[12] A. Richman and P. Schone. Mining Wiki resources for multilingual named entity recognition. In *ACL*, 2008.

[13] P. Schönhofen, A. Benczúr, I. Bíró, and K. Csalogány. Performing cross-language retrieval with Wikipedia. In *CLEF*, 2007.

[14] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *KDD*, 2008.

[15] S. Teufel and H. V. Halteren. Evaluating information content by factoid analysis: Human annotation and stability. In *ACL*, 2004.

[16] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. McGuinness. Computing trust from revision history. In *PST*, 2006.

Ubiquity: Designing a Multilingual Natural Language Interface

Michael Yoshitaka Erlewine
Mozilla Labs
1981 Landings Drive
Mountain View, CA 94043
mitcho@mitcho.com

ABSTRACT

This paper describes the design and implementation of Ubiquity, a multilingual textual interface for the Firefox browser developed at Mozilla Labs. The Ubiquity interface facilitates rapid information retrieval and task execution in the browser, leveraging existing open web APIs. The importance of offering equivalent user experiences for speakers of different languages is reflected in the design of Ubiquity's new natural language parser, described here. This paper also aims to advocate the further development of equipotent multilingual interfaces for information access.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems—*human information processing*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*natural language*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language parsing and understanding*

General Terms

Design, Experimentation, Human factors, Languages

1. INTRODUCTION

Language continues to be one of the greatest barriers to open information access on the internet. While much effort and increased attention have been devoted to the development of multilingual corpora and resources, less attention has been given to guaranteeing that users with different linguistic backgrounds can use the same quality tools to access that information. As part of Mozilla's goal to make the internet experience better for all users [5], Ubiquity aims to bring a new form of interactivity into the browser which treats user input in different languages equally. Ubiquity offers a platform for rapid information access, with no languages treated as second-class citizens.

Ubiquity was borne out of the Humanized Enso product (http://www.humanized.com/enso/), but is now an open-source community project, with dozens of contributors and active testers. It is available for download at http://ubiquity.mozilla.com and can be installed on the Firefox browser. Similar popular text-based command interfaces which are overlaid on GUI include Quicksilver (http://www.blacktree.com) and GNOME Do (http://do.davesbd.com/), but neither of them attempts a natural language syntax, nor do they support localization of their parser and keywords.

2. TOWARDS A NATURAL INTERFACE

2.1 Features of a Natural Syntax

The lead of Ubiquity development Aza Raskin argues in his 2008 ACM *interactions* paper that text-based interfaces can be more humane than overextended graphical interfaces [7].[1] Graphical interfaces are easy to learn and apply for concrete tasks but do not scale well with additional functionality and lack the precision required to communicate abstract instruction. While numerous text-based computer interfaces exist, they have been deemed too difficult for lay users. Raskin argues that textual interaction does not entail these difficulties per se; rather, they are products of their oft-times stilted grammars. In reconsidering the text-based interface, ease and familiarity built into the interface are key. A subset of natural language is thus a clear winner.

Many programming and scripting languages—themselves interfaces to instruct the computer—make use of keywords inspired by natural languages (most often English). Many simple expressions neatly mirror a natural language (1a) but more complex instructions will quickly deviate (1b).

(1) a. `print "Hello World"` (Python)
b. `print map(lambda x: x*2, [1,2,3])`

One valiant effort to facilitate near-natural language instruction has been AppleScript, which enables complex English-like syntax (as in 2) and originally was planned to support similar Japanese and French "dialects."

(2) `print pages 1 thru 5 of document 2` (AppleScript)

[1]The term "humane" is used in this paper to describe human-computer interfaces which are "responsive to human needs and considerate of human frailties" [9] (see also [8]).

As a full-featured scripting language, however, more complex expressions push beyond the natural language metaphor and introduce their own idiosyncrasies. Bill Cook, one of the original developers of AppleScript, notes "in hindsight, it is not clear whether it is easier for novice users to work with a scripting language that resembles natural language, with all its special cases and idiosyncrasies" [2]. Raskin notes that this is precisely what must be addressed in designing a humane text-based interface: "if commands were memorable, and their syntax forgiving, perhaps we wouldn't be so scared to reconsider these interface paradigms" [7].

In designing an internationalizable natural language interface, we can conclude that it is not enough to use natural language keywords and mimic its syntax. The grammar must never conflict with a user's natural intuitions about their own language's syntax—a goal I call *natural syntax*. While a user can't expect such an interface to understand every natural language command, a good rule of thumb is that multiple natural alternatives for a given intent are interpreted in the same way. For example, consider the examples (3) in Japanese, a language with scrambling.[2]

(3) a. 太郎に ボールを 投げろ
Taro-ni ball-o nagero
Taro-DAT ball-ACC throw-IMPER

b. ボールを 太郎に 投げろ
ball-o Taro-ni nagero
ball-ACC Taro-DAT throw-IMPER

Both sentences are valid expressions for the command "throw a ball to Taro." An interface with a natural syntax must understand either both of these inputs or, if for example the interface does not understand the verb *nagero*, neither of them. To understand one but not the other goes against the tenet of natural syntax.

2.2 Commands in Ubiquity

Ubiquity actions are requests for actions or information, corresponding functionally to the formal clause type of "imperative" [6], although they may manifest in forms traditionally characterized as "imperative," "infinitive," or "subjunctive," depending on the language [4]. No vocative is entered as the addressee is always the computer, nor do we handle negation,[3] leaving Ubiquity input to simply be composed of a single verb and its arguments (if any). Some example English Ubiquity actions include:

(4) a. `translate hello to Spanish`—previews the text "hola." On execution, inserts the text "hola" in the active text field.

b. `email hello to John`—on execution, composes a new email to contact `John` with message body "hello."

[2]Note that the Japanese examples are given with spaces between words to facilitate the glosses. Japanese does not normally place spaces between words.

[3]When negative imperative meanings are desired, verbs which lexicalize the negative meaning are chosen, e.g. `prevent`, `turn off`, etc.

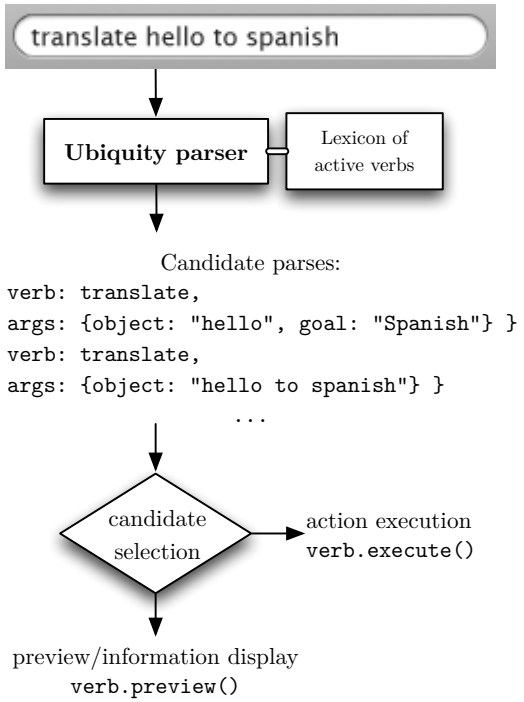


Figure 1: Schematic diagram of user interaction with Ubiquity.

c. `map Tokyo`—previews a map of Tokyo using the Google Maps API. The image can then be inserted into the page.

Verbs are written in JavaScript. Each verb may specify a `preview()` method which displays some information to the user or gives a preview of the action to be executed and an `execute()` method which carries out the intended action.

In order to avoid ambiguity, a list of possible parses is presented to the user for confirmation before execution. Suggestions give a visual indication of the parsing. A scoring mechanism is used to bring more likely candidates to the top, taking user input and browser use habits into consideration.

3. ADDRESSING THE NEEDS OF MULTILINGUAL ACCESS

With the requirements and goals of the project as laid out in section 2, certain architectural choices were made in designing the parser in order to support multiple languages equipotentially. In this section I will review the unique features of our parser and platform which enable equal information access and rapid localization.

3.1 Identifying Arguments by Semantic Role

Ubiquity commands' ease of creation is a great strength for the platform, with many contributors around the world creating and sharing their own verbs as well as writing new verbs for personal use. In order to let users of different languages benefit equally from the platform, however, there is a need to internationalize the verbs themselves. Verbs include

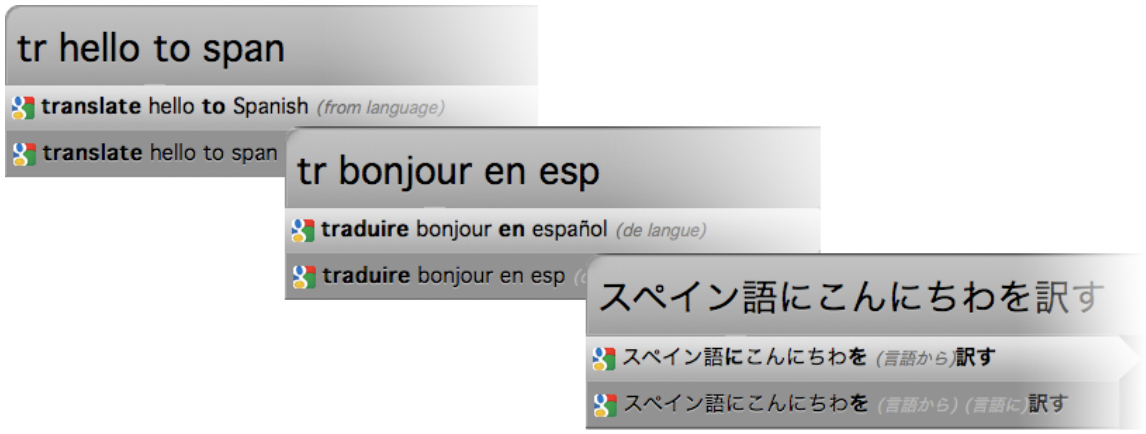


Figure 2: Equivalent Ubiquity queries in three languages: English, French, and Japanese. Note that the two suggestions returned in each case are semantically different, reflecting the ambiguity between translating "hello to span" into an as yet unspecified language and translating "hello" into the Spanish language.

some strings which must be translated, such as the verb's name, but they also include a specification of the type of arguments it accepts, known as the *syntactic frame* of the verb. For example, in English an `email` verb may take a direct object and a recipient introduced by the preposition "to," while a `translate` verb may take an arbitrary direct object, a goal language marked by "to," and a source language marked by "from."

In order to facilitate this localization, we chose to let verbs specify their syntactic frames using abstract semantic roles such as `object`, `goal`, `instrument`, `position`, etc. which are morphosyntactically coded in most languages.[4] For example, suppose an English-speaking contributor wrote a verb called `move`, whose action was to move an object from one location to another. Its syntactic frame could be specified as follows, where `physical_object` and `location` are noun types which specify a class of arguments and their associated semantic forms.

```
{ object: physical_object,
  source: location,
  goal:   location }
```

The command author could then use this command in English, entering input such as (5). The parser recognizes the English prepositions "to" and "from" as corresponding to the `goal` and `source` roles (underlined below), and recognizes the unmarked argument as an `object`.

(5) `move truck from Paris to Beijing`

(6) トラックをパリから北京へ`move`

However, given a set of localized noun types, the exact same command code could be used with the Japanese parser by entering the input (6). Here, the parser recognizes that the postpositions "を," "へ," and "から" mark `object`, `goal`, and `source` arguments, respectively. The only manual localization required for the `move` command, then, is the translation of the verb name "move" itself. As seen by this example, the specification of arguments using abstract semantic roles supports the rapid and, indeed, semi-automatic localization of commands, ensuring that users of all languages benefit from individual improvements to the Ubiquity platform's functionality.

[4]In our use, "semantic role" is equivalent to the related notions of "grammatical function" and "thematic relation." An inventory was chosen based on [3] and subsequent cross-linguistic work.

Table 1: Argument-first Suggestions

Sample argument parses	suggested verbs
`{ object:…, goal:… }`	email, send
`{ object:…, instrument:… }`	search, look up
`{ object:…, source:…, goal:… }`	move, translate

3.2 Argument-first Suggestions

In parsing Ubiquity input, a key task is the identification of the verb, if any. In many languages the verb naturally comes at the beginning of the sentence (see English examples in 4). In this case, as the verb can be identified early in the user input, we can then annotate the candidate parses with information on the missing arguments to guide the user in entering the rest of their input (see figure 2). However, not all languages enter the verb first in commands. Some languages are strictly verb-final (e.g. Japanese), while in some other languages (e.g. German, Dutch, Modern Greek) it is equally valid to express commands using the imperative or subjunctive verb form at the beginning of the sentence, or using the infinitive at the end of the sentence.

Rather than being discouraged by this conundrum, thought was given to how we can leverage the unique qualities of verb-final (or argument-first) input to make a more humane and supportive interface. As different verbs in our lexicon specify different syntactic frames, by parsing arguments and identifying semantic roles in the input before the verb is known, we can then suggest verbs to the user which match that particular argument structure (see examples in table 1 of some such suggestions). This smart argument-first suggestion aids in command discoverability by suggesting verbs for a given target which the user may not have known

existed. This approach crucially takes advantage of the argument-first input and offers unique value and increased usability to users with verb-final languages.

Note also that the suggestion of verbs based on argument-only input may also be useful for regularly verb-initial langauges such as English. Studies of general interactive systems concur that *noun-verb* (or *object-action*) paradigms result in error reduction, increased speed, and better reversibility during input [8]. For these reasons, argument-first suggestions are supported in Ubiquity for all languages equally.

3.3 Minimal Language Descriptions

The Ubiquity parser attempts to make as much of its parser algorithm universal as is practical, taking a page from the Principles and Parameters framework in generative linguistics.[5] A single universal parser was designed, with settings for different languages built on top of that base [1]. The settings for each language are written in JavaScript and range from ten to thirty lines of code. Various hooks exist in the code for language-specific processing when necessary, but the majority of the language settings are simply lists of special lexical items such as the prepositions or postpositions in a language. In this way, for the limited range of data which constitute Ubiquity input, the very difficult problem of writing a language-specific parser is reduced to little more than some native speaker consultation and string translation.

4. CONCLUSIONS

Further globalization of the web without serious consideration of multilingual information access could spur the further fragmentation of information and ideas. Equal access to information will require more than just cross-language search and retrieval systems, but also universal interfaces which are designed for rapid localization and treat all languages equally.

In this paper I outlined some of the design features of Ubiquity's interface and natural language parser which bring the system closer to this goal. Formal approaches to the study of language were applied in order to design a system which can be extended to a wide range of languages. As of this writing, settings for nine languages have been written for Ubiquity, while the community process of setting technical standards for verb and noun type localization has just begun.

Ubiquity is quickly becoming a compelling text-based interface for both advanced and casual users. The forgiving "natural syntax" philosophy and the smart suggestion of verbs and arguments to the user help make Ubiquity a humane interface which cooperates with users rather than confounds them. These qualities make Ubiquity a natural choice of interface platform for multilingual and cross-languge information access applications.

5. ACKNOWLEDGMENTS

Thank you to comments from Aza Raskin and Jonathan DiCarlo at Mozilla and audiences at BarCamp Tokyo; Tokyo 2.0; Chuo, Waseda, and Keio Universities; as well as comments on related material on my blog (http://mitcho.com/blog/).

[5]It is worth noting that this architectural choice also complemented the object-oriented architecture and Don't Repeat Yourself design goals of the project.

6. REFERENCES

[1] Parser 2 - MozillaWiki. https://wiki.mozilla.org/Labs/Ubiquity/Parser_2 .

[2] W. R. Cook. Applescript. In *The Third Conference on the History of Programming Languages*, 2007.

[3] C. J. Fillmore. *Types of lexical information*. Reidel, Dordrecht, 1969.

[4] S. Iatridou. De modo imperativo. Lecture notes, ENS, Paris, September 2008.

[5] Mozilla Foundation. The Mozilla manifesto, v0.9. http://www.mozilla.org/about/manifesto.en.html .

[6] P. Portner. The semantics of imperatives within a theory of clause types. In *Proceedings of Semantics and Linguistic Theory*, volume 14, 2005.

[7] A. Raskin. The linguistic command line. *interactions*, 15(1):19–22, 2008.

[8] J. Raskin. *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, 2004.

[9] A. Varma. Humanized > Why 'humane' is a better word than 'usable'. http://humanized.com/weblog/2006/06/01/why_humane_is_a_better_word_than_usable/.

NSContrast: An Exploratory News Article Analysis System that Characterizes the Differences between News Sites

ABSTRACT

The News Site Contrast (NSContrast) system analyzes multiple news sites using the concept of contrast set mining and can extract the terms that characterize the differences in topics of interest for each country. However, because of the poor quality of some machine translation, NSContrast results include some meaningless terms generated by this mistranslation. To address this problem, Wikipedia is used as a bilingual dictionary and as a source for synonym identification. We give some experimental results for this New NSContrast system.

Categories and Subject Descriptors

H3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Information Retrieval, Text Mining

Keywords

News search, IR interface, Contrast set mining

1. INTRODUCTION

It has recently become possible to access a wide variety of news sites from across the world via the Internet. Because each news site has its own culture and interpretation of events, we can obtain a greater diversity of information using multiple news sites than ever before.

Because each country has different opinions and interests, when we use news sites from different countries, we will obtain different points of view for a topic. For example, considering diplomatic issues to do with North Korea; Asian, European and American news sites have some common interests and their own characteristic interests. Therefore, to analyze events reported from multiple sites, it is important to clarify the characteristics of each news site.

Boston 2009 USA

There are several experimental systems that integrate news articles about a particular event from multiple news sites. For example, EMM News Explorer [1] and Newsblaster [4] are integrated news aggregation systems from distributed news archives.

These systems are effective for understanding a particular event via multiple news sites, but they ignore the characteristics of each news site. For example, Japanese news sites tend to report Japanese-related topics more frequently than others. To better understand articles from different news sites, this bias should be taken into account. To identify news site characteristics, NSContrast [9] has been proposed This system analyzes multiple news sites using the concept of contrast set mining and aims to extract the characteristic information about each news site by performing term co-occurrence analysis. The system has potential for extracting characteristic terms that reflect topic divergence between different countries. However, because of the poor quality of some machine translation, NSContrast results include some meaningless terms generated by this mistranslation [8].

In this paper, we discuss the issues related to handling articles from multiple news sites via machine translation, and we propose a method for using Wikipedia as a resource for constructing a bilingual dictionary for an NSContrast system. Using this dictionary, a new database of news articles is constructed, which is then tested via user experimentation with an NSContrast system.

2. NSCONTRAST

2.1 Term Collocation Analysis by Contrast Set Mining

Term collocation analysis is a well-known text mining method for extracting characteristic information from texts[6]. However, conventional collocation analysis mostly focuses on the characteristic information that is dominant in the text database. In many cases, most of the information is well known and is therefore not particularly interesting.

To solve this problem, we introduce the concept of contrast set mining [7] for the analysis. This framework compares a global data set and a conditioned local data set to find characteristic item information that is significantly different from the global characteristic information. Even though this information may not be dominant in either the global or the local data set, it can be used to understand the characteristics of the local database.

[1] http://press.jrc.it/NewsExplorer/

We use Discovery of Correlation (DC) pair mining [7] for term collocation analysis. In DC pair mining, the "difference in correlations observed by conditioning a local database" is of particular interest. To quantify this difference, we introduce a new measure, $change(X, Y; C)$, defined by

$$change(X, Y; C) = \frac{correl_C(X, Y)}{correl(X, Y)},$$

where X and Y represent the item sets and C represents the condition for creating the local database. $correl(X, Y)$ and $correl_C(X, Y)$ correspond to the correlations between X and Y in the global database and a C-conditioned local database, respectively.

To utilize this technique for term collocation analysis of multiple text databases (news text databases from different countries), we modify this analysis method as follows.

- Because the size of each text database is not much smaller than that of the global database, a characteristic collocation occurring only in one database is also a characteristic collocation in the global database. Therefore, the contrast between the conditioned database and the rest of the databases is used instead of the original definition.
- Because the computational cost of DC pair mining is substantial, the target term for analysis (X in the formula) is given by the user.

By using this technique, we can extract characteristic minor topics that are of interest: higher change (X_m in Figure 1) or are neglected: lower change (X_n in Figure 1) in one database compared with others.

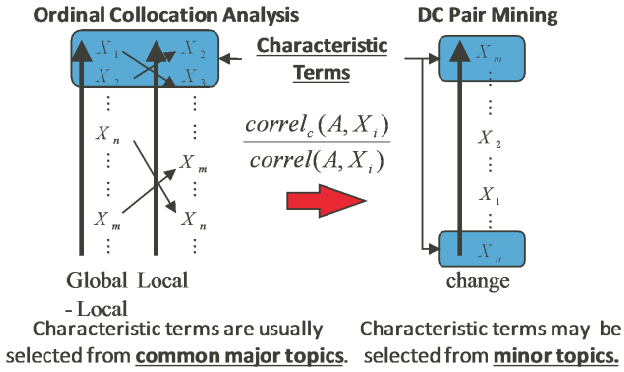


Figure 1: Collocation Analysis based on DCPair Mining

2.2 NSContrast: A News Site Analysis System

The NSContrast system is a method for accessing news articles from multiple news sites using the concept of DC pair mining[9, 8]. This system has the following analytic components.

- Term collocation analysis based on DC pair mining.
  The system generates a list of characteristic terms by comparing news article databases from different countries. This term list is represented as a term collocation graph to aid understanding of the relationships among characteristic terms.
- A burst analysis function [3] for finding an appropriate time sequence window.
  To find good characteristic terms using contrast set mining, it is preferable to select a large number of articles for a particular topic. Because burst analysis is a method that finds a period during which a given term is of more interest than usual, it is an effective technique for finding this information.
- A news article retrieval system.
  To understand the meaning of term collocation analysis and burst analysis, a news article retrieval system is used for this purpose.

3. A NEW NSCONTRAST USING WIKIPEDIA INFORMATION

We constructed a news article database by collecting news articles from news sites in Japan, China, Korea, and the USA. We confirmed that NSContrast has capabilities that extract characteristic terms for understanding the differences between news articles on a particular topic from multiple countries.

However, during experiments, a user claimed that the system tends to select the following two types of meaningless characteristic terms.

1. Terms generated by mistranslation.
   For example, "奥巴馬" means Obama (President of the United Sates) in Chinese, but the machine translation system translates this word as "オーストラリア巴馬". ("奥" is also used for representing "オーストラリア (Australia)"). Because of this mistranslation, NSContrast tends to select "オーストラリア (Australia)" as a characteristic term for China when the system is analyzing topics related to the President of the United States.
2. Terms with many synonyms.
   For example, "北朝鮮 (North Korea)", "朝鮮民主主義人民共和国 (Democratic People's Republic of Korea)", and "DPRK" are equivalent terms for the same country in Asia. When one of the news sites uses a different representation from others, this term tends to be selected as a characteristic term.

From an analysis of these problems, we found that they occur mainly when handling named entities (e.g., names of people, countries, and companies) in a Chinese–Japanese translation system. This is because the dictionary for the system is updated only infrequently and therefore lacks entries for recently named entities.

Because Wikipedia has many entries related to named entities and these entries are associated via language links (equivalent relationships among different languages) and redirection links (reference relationships in the same language), it is a good resource for constructing both bilingual dictionaries and synonym dictionaries.

In this research, we constructed a Chinese–Japanese bilingual dictionary from Wikipedia, based on the method proposed by [2]. Because it is easier for the machine translation system to translate Japanese named entities that are described in terms of Chinese characters than katakana's one

(katakana contains phonograms that are used mainly to represent words imported from other countries), we constructed a bilingual dictionary for Japanese katakana terms, as follows.

1. Selection of Japanese katakana entries with language links.
   From a Japanese Wikipedia data dump, we selected Japanese katakana entries that have a language link to Chinese entries.
   For example, "バラク・オバマ (Barack Obama)" is selected because it is a katakana entry and has a Chinese language link to "巴拉克・奥巴馬".

2. Separation of first name and family name.
   For most people, in both katakana and Chinese, the first name and the family name are separated by "・". The correspondence between these names is specified by using "・".
   For example, from a language link "バラク・オバマ (Barack Obama)" to "巴拉克・奥巴馬", two dictionary entries ("巴拉克" to "バラク (Barack)" and "奥巴馬" to "オバマ (Obama)") are generated.

3. Use of redirection links in Chinese Wikipedia.
   In the Chinese language, two or more different Chinese expressions for the same person often exist. In such cases, redirection links in the Chinese Wikipedia are used to represent the relationship among these entries. By using these links, bilingual dictionary entries for handling these terms were generated.
   For example, from a Chinese redirection link "歐巴馬" to "奥巴馬" and a language link "奥巴馬" to "オバマ (Obama)", a new dictionary entry "歐巴馬" to "オバマ (Obama)" is generated.

By using this method, 13,255 bilingual dictionary entries (Chinese to Japanese) were extracted from Japanese and Chinese Wikipedia data dumps.

We used this Chinese–Japanese dictionary as an add-on external bilingual dictionary for a machine translation system by using the Language grid service [1, 5] that supports the combination of a dictionary lookup service with a machine translation system.

When we used this machine translation system with the Wikipedia dictionary, we found several entries that were inappropriate translations. For example, from the Wikipedia entry "フレンズ (Friends: TV drama from the US)", dictionary entry "朋友" to "フレンズ (Friends)" was generated. However, because "朋友" is also used as a common word (i.e., an unnamed entity word), it is preferable to use "友人 (Friends)" instead of "フレンズ (Friends)" in such cases.

We checked manually the frequently occurring dictionary entries in a Chinese newspaper database to remove such inappropriate entries.

We also used Japanese redirection links to construct a synonym dictionary. All entries that had redirection links were normalized with respect to their destination entries. For example, using the redirection link "北朝鮮 (North Korea)" to "朝鮮民主主義人民共和国 (Democratic People's Republic of Korea)", "北朝鮮 (North Korea)" was normalized to "朝鮮民主主義人民共和国 (Democratic People's Republic of Korea)" in the database.

In addition, because Wikipedia has a broad coverage of entries, from general topic terms to technical terms and from traditional terms to newly introduced terms, these entries may be good candidates for characteristic terms.

Based on this understanding, we developed a New NSContrast system by modifying the NSContrast database to include an index of Wikipedia entries for each news article. We also restricted the candidates for term collocation analysis to Wikipedia entries.

The news article database was populated by following news sites from January 1, 2008 to April 19, 2009. Table 1 shows detailed information of the database.

Table 1: News Site Information

Name of the site(country) URL	Number of articles Total/Daily
Asahi newspaper (Japan) http://www.asahi.com/	58344/122
Yomiuri newspaper (Japan) http://www.yomiuri.co.jp/	48675/102
Nihon keizai newspaper (Japan) http://www.nikkei.co.jp/	69638/146
CNN (USA) http://www.cnn.co.jp/	9542/20
Chosun newspaper (Korea) http://www.chosunonline.com/	24389/51
Joins newspaper (Korea) http://japanese.joins.com/	18842/39
People newspaper (China) http://j.peopledaily.com.cn/	18775/38
Chosun newspaper (Korea) Machine Translation http://www.chosun.com/	102009/214
Xinhua newspaper(Chine) Machine Translation http://www.xinhuanet.com/	367459/773

Figure 2 shows an example of New NSContrast output when analyzing the term "Pakistan" during the first half of December, 2008. All terms were originally in Japanese and are translated into English. Blue lines and comments are added manually as a summary of the news article retrieval based on the related terms. Comparing these results with those for the previous NSContrast system, the number of articles that have "Pakistan" as an index term has increased because of the synonym assimilation. In addition, the readability of the word collocation graph has improved.

4. CONCLUSION

In this paper, we have introduced the NSContrast system, which can extract characteristic information about news sites for given topic terms by using contrast set mining techniques. In addition, we have also proposed using Wikipedia to handle different expressions of the same named entity. By using Wikipedia in a New NSContrast system, we confirmed an improved recall of the news article retrieval results for a particular topic. This is important for an unbiased analysis of the differences between various countries.

5. REFERENCES

[1] A. Bramantoro, M. Tanaka, Y. Murakami, U. Schäfer, and T. Ishida. A hybrid integrated architecture for

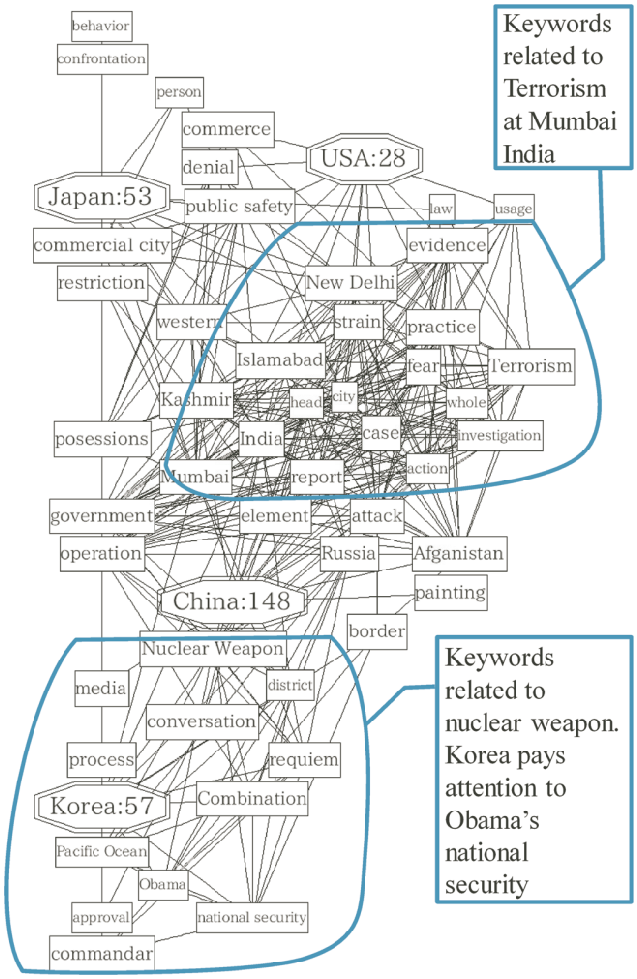


Figure 2: Term Collocation Graph for the Term "Pakistan"

language service composition. In *ICWS '08: Proceedings of the 2008 IEEE International Conference on Web Services*, pages 345–352, Washington, DC, USA, 2008. IEEE Computer Society.

[2] M. Erdmann, K. N. T. Hara, and S. Nishio. An approach for extracting bilingual terminology from wikipedia. In *Proc. of International Conference on Database Systems for Advanced Applications (DASFAA) (Mar. 2008)*, pages 686–689. Springer-Verlag GmbH, 2008. LNCS 4947.

[3] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 91–101, New York, NY, USA, 2002. ACM Press.

[4] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*, 2002.

[5] NICT Language Grid Project. *How to use Language Service*, 2009. http://langrid.nict.go.jp/langrid-developers-wiki-files/language-service-use-manual-20090219-en.pdf.

[6] F. Smadja. Retrieving collocations from text: Xtract. *Comput. Linguist.*, 19(1):143–177, 1993.

[7] T. Taniguchi and M. Haraguchi. Discovery of hidden correlations in a local transaction database based on differences of correlations. *Data Engineering Applications of Artificial Intelligence*, 19(4):419–428, 2006.

[8] M. Yoshioka. Analyzing multiple news sites by contrasting articles. In *Proceedings of the Fourth Intl. Conf. on Signal-Image Technology & Internet-Based Systems*, pages 45–51, 2008.

[9] M. Yoshioka. Ir interface for contrasting multiple news sites. In H. Lee, T. Liu, W.-Y. Ma, T. Sakai, and Kam-FaiWong, editors, *Information Retrieval Technology 4th Asia Infomation Retrieval Symposium, AIRS 2008, Harbin, China, January 15-18, 2008 Revised Selected Papers*, pages 508–513. Springer-Verlag GmbH, 2008. LNCS4993.

Multilingual Ontologies for Information Access

Elena Montiel-Ponsoda
Ontology Engineering Group,
Dpto. de Inteligencia Artificial
Facultad de Informática, UPM
28660 Boadilla del Monte, Madrid,
Spain. Tel: +34 91 336 3670
emontiel@fi.upm.es

Mauricio Espinoza
Ontology Engineering Group,
Dpto. de Inteligencia Artificial
Facultad de Informática, UPM
28660 Boadilla del Monte, Madrid,
Spain. Tel: +34 91 336 3670
jespinoza@fi.upm.es

Guadalupe Aguado de Cea*
Ontology Engineering Group,
Dpto. de Inteligencia Artificial
Facultad de Informática, UPM
28660 Boadilla del Monte, Madrid,
Spain. Tel: +34 91 336 3670
lupe@fi.upm.es

ABSTRACT

For most international organizations multilinguality is a crucial issue, since they have to guarantee users all over the world an equal access to information. Not only access to data in different natural languages is vital, but also the annotation and indexing of that multilingual information for its management within organizations. We deem multilingual ontologies to provide a viable solution to the need for accessing and managing multilingual information, since they can organize great amounts of data, and offer sound and powerful means for reason with it. However, the availability of multilingual ontologies is still very limited, and tools for localizing ontologies, as well as methods for supporting the inclusion of multilingualism in ontologies are urgently demanded. With our research, we aim at approaching both aspects of multilinguality in ontologies. On the one hand, we present a tool for an automatic localization of ontologies from a source to a target natural language. On the other hand, we propose a model for the association of multilingual information to ontologies.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms

Design, Languages

Keywords

Ontology localization, multilingual ontologies, LabelTranslator, Linguistic Information Repository (LIR)

1. INTRODUCTION

International organizations working in a multilingual environment are becoming more and more conscious of the impending need for resources that allow them to manage the huge amounts of data and linguistic resources they have to deal with in different natural languages. Moreover, these organizations have to guarantee effective information access to users speaking different languages. In this sense, ontologies are sound and powerful resources that would come to solve this need. However, the multilingual information ontologies can nowadays support falls short of meeting international organization requirements. Finally, the effort implied in the construction of multilingual ontologies makes them nearly inexistence in the current Web.

With the aim of taking advantage of the substantial amount of ontologies currently available on the Web in one natural language, ontologies have to undergo *localization*, understood as the adaptation of existing ontologies to a concrete language and culture community [[12]]. To automate this activity we propose LabelTranslator, a system that takes as input an ontology whose labels are described in a source natural language and obtains the most probable translation of each label in a target natural language. The automation of this process would reduce the human efforts implied in the manual localization of ontologies. For this aim, LabelTranslator relies on available translation services and multilingual resources, and sorts out translation senses according to similarity with the lexical and semantic context of each ontology label.

Then, the information obtained from the localization process by LabelTranslator is stored in the so-called Linguistic Information Repository or LIR, an external portable model that permits to associate rich linguistic information to any element in the ontology. In particular, the LIR contributes to the localization of the ontology terminological layer, i.e. terms or labels that name ontology elements. The main features of the LIR are related with: (1) the establishment of well-defined relations of lexicalizations within and across languages; (2) the representation of conceptualization mismatches among different cultures and languages; and (3) a unified access to aggregated multilingual data.

The rest of this paper is structured as follows. In section 2 we give an overview of the main functionalities of LabelTranslator, as well as some references to related work. Then, in section 3 we summarize the main ways we have identified to obtain a multilingual system. Section 4 is devoted to an outline of the Linguistic Information Repository. Finally, the paper is concluded in section 5.

Conference'04, Month 1–2, 2004, City, State, Country.

2. LABELTRANSLATOR FUCTIONAL OVERVIEW

LabelTranslator [4][5][6] has been designed with the aim of automating ontology localization, and has been implemented in the ontology editor NeOn Toolkit[1] as a plug-in. In its current version, it can localize ontologies in English, German and Spanish. In the following, we briefly describe the main tasks followed by the system in performing the localization activity.

Once an ontology has been created or imported in the NeOn ontology editor, LabelTranslator allows users and domain experts to manually sort out the ontology terms that should undergo localization. By default the system selects the whole ontology. For each ontology element, LabelTranslator retrieves its *local context* (set of hypernyms, hyponyms, attributes, and sibling labels associated with the ontology term under consideration), which is interpreted by the system using a *structure-level* approach.

In order to obtain the most appropriate translation for each ontology term in the target language, LabelTranslator accesses multilingual linguistic resources (EuroWordNet[2], Wiktionary[3], or IATE[4]) and translation web services (GoogleTranslate[5], BabelFish[6], etc.) available on the Web. From these resources, the system obtains equivalent translations for all selected labels. Then, it retrieves a list of semantic senses for each translated label, querying remote lexical resources as EuroWordnet or third-party knowledge pools such as Watson[7], which indexes ontologies available on the Web. Finally, the senses of each context label are as well discovered following the strategy just explained. At this point, it should be noted that LabelTranslator includes a compositional method to translate compound labels, which first searches for translation candidates of each token of the compound label, and then builds the translations for the candidates using lexical templates. For a detailed explanation see [5].Then, the system uses a disambiguation method to sort out the translations according to their context. LabelTranslator carries out this task in relation to the senses of each translated label and the senses of the context labels. At this stage, domain and linguist experts may decide to choose the most appropriate translation from the ones in the ranking. In default, the system will consider the one in the highest position.

The ontology is updated with the resulting linguistic data, which is stored in the LIR model, as will be explained in more detail in section 3. In Figure 1 a screenshot of the LIR API in the NeOn Toolkit is presented.

2.1 Related Work

Our work enhances the work presented in [3], where a system for supporting the multilingual extension of ontologies expressed in just one natural language was proposed. This tool was used to support “the supervised translation of ontology labels". Therefore, the tool offers a semi-automatic strategy. In our approach we have implemented an automatic method to reduce human intervention while enriching an ontology with linguistic information.

In [10] the authors propose a framework for adding linguistic expressivity to conceptual knowledge, as represented in ontologies. They use two lexical resources for linguistic or multilingual enrichment: WordNet, and DICT dictionaries. In this work, the process to translate compound ontology labels is not described.

In [7] a method to give support to multilingual ontology engineering is developed. In this work some software tools have been used for supporting the process of term extraction and translation. In particular, the translation process requires sentence aligned parallel text, which has been previously tokenized, tagged and lemmatized. In our opinion, obtaining an aligned corpus is not a simple task. Unlike this work, we rely on some multilingual translation services and extend them using lexical templates.

3. REPRESENTING MULTILINGUALITY IN ONTOLOGIES

Regarding the activity of Ontology Localization, we have identified three ways of modelling multilinguality in ontologies:

1) inclusion of multilingual labels in the ontology by means of the `rdfs:label` and `rdfs:comment` properties (most widespread modality)
2) mapping of several conceptualizations in different natural languages through an interlingual set of common concepts (as in the well-known EWN[8] lexicon)
3) association of an external linguistic model to the ontology (as in LexInfo [1] or LingInfo [2]).

The first modality option restricts the amount and type of linguistic information that can be associated to the ontology. The second option requires a huge effort at two stages: first, when a new language has to be integrated in the multilingual system, since a new conceptualization has to be developed, and second, by the establishment of alignments among conceptualizations or between the new conceptualization and the interlingua. In this way, our hypothesis is that the best solution lies on the third option, in which the type and quantity of linguistic information is not restricted, and the linguistic elements that compose the model can be related among them. Regarding this latter option, we argue that existing models have not been intended to cover localization needs, and do not include enough information in this sense, but rather focus on other linguistic information such as the morphosyntactic realizations of ontology labels (see [2]), or their subcategorization frames (the syntactic arguments activated by a word in a sentence, see [1]).

[1] http://www.neon-toolkit.org

[2] http://www.illc.uva.nl/EuroWordNet

[3] http://en.wiktionary.org/wiki

[4] http://iate.europa.eu

[5] http://www.google.com/translate_t

[6] http://babel¯sh.altavista.com

[7] http://watson.kmi.open.ac.uk/WatsonWUI

[8] http://www.illc.uva.nl/EuroWordNet

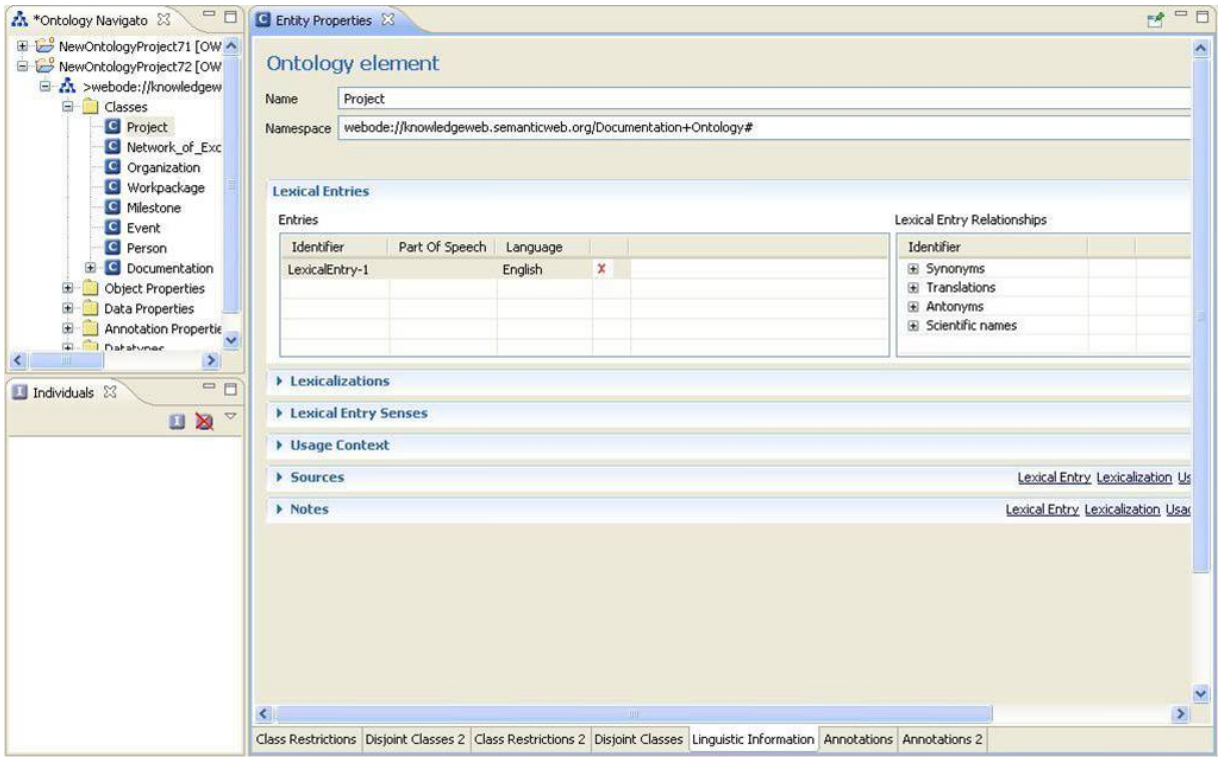


Figure 1. *Linguistic Information* Entity Properties View in NeOn Toolkit

.

4. LINGUISTIC INFORMATION REPOSITORY (LIR)

With the aim of providing available ontologies in one natural language with multilingual information that contributes to their localization, in the NeOn project we have designed the LIR [8][9][11], an external linguistic model based on existing linguistic (LMF[9]) and terminological (TMF[10]) representation schemas. The LIR permits the association of a set of linguistic data to any element in the ontology. The main classes or data categories that compose the LIR are: *LexicalEntry*, *Lexicalization*, *Sense*, *Definition*, *Language*, *Source*, *Note*, and *Usage Context* (as can be seen in Figure 2). Thanks to the relations that can be established among the LIR classes, the LIR mainly accounts for: well-defined relations within lexicalizations in one language and across languages, and conceptualization mismatches among different cultures and languages. The main benefits of this approach against the modeling options presented in section 2 are: a) the association of an unrestricted quantity of linguistic information to ontology elements; b) the establishment of relations among the linguistic elements, as well as the performance of complex operations (reasoning) with them; c) the access and manipulation of the linguistic data (terminological layer) without interfering with the conceptualization, with the resulting benefits for non-ontology engineers; and d) the reuse of the contained linguistic information for other applications.

Up to now, the LIR has been implemented as an ontology in OWL [5], and is supported by the LabelTranslator NeOn plug-in, as mentioned in section 2. A first set of tests has been conducted within NeOn to asses the suitability of the LIR model for the linguistic needs of the Food and Agriculture Organization of the United Nations (FAO). The LIR has proved to satisfy the FAO needs for i) establishing relations among lexicalizations within and across languages, ii) specifying variants for dialects or local languages, and iii) explicitly expressing translation specificities.

5. CONCLUSIONS

In this paper our objective was to present some technological and modeling solutions for the localization of ontologies. We believe that multilingual ontologies may contribute a reliable access and management of great amounts of multilingual data. For this aim, we provide LabelTranslator, a tool for an automatic localization of ontologies from a source natural language to a target natural language. The information obtained from the localization activity with LabelTranslator is stored in an external portable model called Linguistic Information Repository or LIR. The LIR allows including a wide range of linguistic elements that account for linguistic variants within the same language and cultural differences among languages. The LIR is currently being aligned with other linguistic representation models and standards to assure interoperability. Evaluation is also being conducted to check the suitability of its present set of classes and relations. Enhancements to the present model are foreseen to better capture relations among and across language.

[9] Lexical Markup Framework ISO/CD 24613

[10] Terminological Markup Framework ISO 16642

With regard to the LabelTranslator NeOn plug-in, its functionalities are being enhanced to be employed in a distributed environment, so that it can be used collaboratively in a scenario where its users are scattered in several locations.

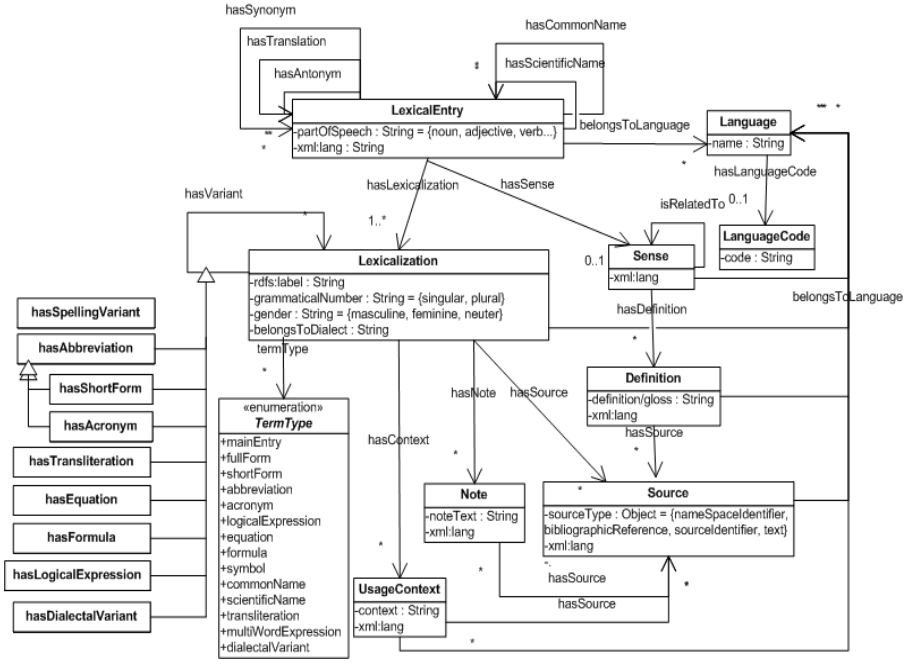


Figure 2. Linguistic Information Repository (LIR)

6. ACKNOWLEDGMENTS

This work is supported by the European Commission's Sixth Framework Program under the project name: *Lifecycle support for networked ontologies* (*NeOn*) (FP6-027595), and the National Project *GeoBuddies* (TSI-2007-65677C02).

7. AUTHORS*

For space reasons, the three remaining authors have been included here:

Asunción Gómez-Pérez, Ontology Engineering Group, Dpto. De Inteligencia Artificial, Facultad de Informática, UPM, 28660 Boadilla del Monte, Madrid, Spain. asun@fi.upm.es. Tel: +34 91 336 3670

Eduardo Mena, IIS Department, Universidad de Zaragoza, María de Luna 1, 50018 Zaragoza, Spain. emena@unizar.es. Tel: +34 976 76 23 40

Wim Peters, Sheffield Natural Language Processing Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK. w.peters@dcs.shef.ac.uk Tel: +44 (0)114 222 1814

8. REFERENCES

[1] Buitelaar, P., Cimiano, P., Haase, P. and Sintek, M. 2009. Towards Linguistically Grounded Ontologies. In Proceedings of the 6th European Semantic Web Conference (ESWC'09), Heraklion, Greece.

[2] Buitelaar, P., M. Sintek, M. Kiesel. 2006. A Multi-lingual/Multimedia Lexicon Model for Ontologies. In Proceedings of the 3rd European Semantic Web Conference (ESWC'06), Budva, Montenegro.

[3] Declerck, T., Gómez-Pérez, A., Vela, O., Gantner, Z. and Manzano-Macho, D. 2006. Multilingual lexical semantic resources for ontology translation. In Proceedings of LREC2006.

[4] Espinoza, M., Gómez-Pérez, A. and Montiel-Ponsoda, E. 2009. Multilingual and Localization Support for Ontologies. In Proceedings of the 6th European Semantic Web Conference (ESWC'09), Heraklion, Greece.

[5] Espinoza, M., Gómez-Pérez, A., Mena, E. 2008. Enriching an Ontology with Multilingual Information. In Proceedings of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain), LNCS Springer, pp. 333-347.

[6] Espinoza, M., Gómez-Pérez, A. and Mena, E. 2008. LabelTranslator - A Tool to Automatically Localize an Ontology. In Proceedings of 5th European Semantic Web Conference (ESWC'08), Tenerife (Spain), LNCS Springer, pp. 792-796, demo paper.

[7] Kerremans, K., and Temmermann, R. 2004. Towards multilingual, termontological support in ontology engineering. In Proceedings Workshop on Terminology, Ontology and Knowledge representation, pp. 22-23.

[8] Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. 2008. Modelling multilinguality in ontologies. In Coling 2008: Companion volume - Posters and Demonstrations, Manchester, UK.

[9] Montiel-Ponsoda, E. and Peters, W. (coordinators) 2008. Multilingual and Localization Support for Ontologies. NeOn Project Deliverable 2.4.2.

[10] Pazienza M.T., and Stellato, A. 2006. Exploiting linguistic resources for building linguistically motivated ontologies in the semantic web. In Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006.

[11] Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G. 2007. *Localizing Ontologies in OWL*. In Proceedings of the OntoLex Workshop, held jointly with ISWC'07, Busan, South Corea.

[12] Suárez-Figueroa, M.C. and Gómez-Pérez, A. 2008. First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008), Copenhagen.

Towards an Integrative Approach to Cross-Language Information Access for Digital Libraries

Jiangping Chen and Miguel Ruiz
Department of Library and Information Sciences, College of Information
University of North Texas
{Jiangping.Chen, Miguel.Ruiz}@unt.edu

1. Introduction

Digital libraries contain human intellectual properties that are valuable to not only its designated user communities, but also to interested organizations and individuals all over the world. The content in most digital libraries are not visible to Web search engines, therefore sophisticated digital library (DL) services such as searching and browsing have been developed and explored to facilitate information access. However, most existing digital libraries can only be accessed in single language. Cross-Language Information Access (CLIA), an extension of the field of Cross-Language Information Retrieval (CLIR) (Jones et al, 2001), enables DL users to search and use digital objects in languages other than their search terms. Although CLIA technologies including CLIR, Cross-language Question Answering (CLQA), Cross-Language Information Extraction (CLIE) has been actively explored by researchers more than decades, none of the technologies have been applied to existing digital libraries to help digital library users access information across languages.

Due to the fact that machine translation usually produce hard-to understand translations, Many organizations and information systems still rely on human translators for translating documents or files from one language to other languages. As for digital libraries, very few digital libraries have realized multilingual information access (Chen, 2007). An analysis of 150 digital libraries found out that only five of them can be accessed by using more than one language. Table 1 lists the five digital libraries.

Table 1. Digital Libraries with Multilingual Information Access

Library Name	URL	Languages
Meeting of Frontiers	http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html	English/Russian
France in America	http://international.loc.gov/intldl/fiahtml/fiahome.html	English/French
Parallel Histories	http://international.loc.gov/intldl/eshtml/	English/Spanish
International Children's Digital Library	http://www.icdlbooks.org/	Digital Objects in 11 languages. Users can do the keyword search in 51 languages.
The Perseus Digital Library	http://www.perseus.tufts.edu	Greek, English, Latin

The above digital libraries have been funded by various funding agencies, especially from the federal government. They are the products of collaboration. DL developers in different countries work in collaboration to produce the bilingual or multilingual

collections. These digital libraries serve broader or global user communities where users speak different languages. However, none of them employ cross-language search or any cross-language information retrieval techniques or machine translation. This situation may be the result of concerns involving technical, financial, and user related issues. However, a lack of exploration on CLIA framework for digital libraries and practical assessment of such framework probably prevents digital library communities from implementing CLIA services.

The European Library (TEL) is one of the projects that has been interested in supporting cross lingual access to the collections of National Libraries members of the EU (Cousins, 2006) but so far it only allows monolingual retrieval in many of the different languages and multilingual access through matching of the controlled vocabulary assigned as part of the metadata. According to Cousins (2006) TEL conducted a survey and found that users have high expectations of the integrated portal to able to support cross language retrieval access. Another recent example of an attempt to create a library with CLIA support is the Europeana project.

On May 23, 2007, Google launched its "Translated Search" in its Google Language Tools (http://www.google.com/language_tools) in addition to other language support services and tools. The launch of the cross-language search by Google was a breakthrough event because it signified the transition from CLIR research to its real application. It was the first time that CLIR and machine translation (MT) were integrated to provide a real application on the Internet. We believe it is time to bring CLIA services to be part of the development of digital libraries.

2. The Integrative CLIA Framework

We are in the position to develop and evaluate an integrative CLIA (iCLIA) framework for digital libraries. The framework will extend the Lexical Knowledge Base (LKB) Model for English-Chinese Cross-Language Information Retrieval (Chen, 2006) to facilitate fast development of cross-language information access (CLIA) services for digital libraries. It also includes procedures and design principles for sustaining the CLIA services.

This framework will have following characteristics:

- Integrating translational knowledge from multiple sources. Translational knowledge is crucial for MT systems and query translation process. The incomplete or faulty translational knowledge is the major cause of the out-of-vocabulary problem. In the iCLIA framework, translational knowledge should be obtained from multiple sources, such as commercial or open-source translation dictionaries, the DL collection, the Web, and the DL users. Efforts will be made to integrate translational knowledge from different resources to develop a lexical knowledge base (LKB) that can be used for translating queries and to augment the dictionary of MT systems for translating documents. Figure 1 depicts the integrative LKB construction process. The simplified structure of the LKB is showed in Figure 2.

- Integrating computational power and human intelligence. CLIA users are usually missing from CLIA literature. Our iCLIA framework emphasizes the interaction with DL users in order to improve the CLIA service for digital libraries. A typical CLIA process for a digital library is represented in Figure 3.
- Integrating various translational disambiguation approaches for best performance. Translation disambiguation has been extensively explored in CLIR literature. Algorithms have been developed for CLIR experiments at IR evaluation forums such as TREC, CLEF, and NTCIR. However, no agreement on single best approach has been achieved yet. IR approach of combining results from multiple solutions maybe interesting to explore for CLIA for DLs.

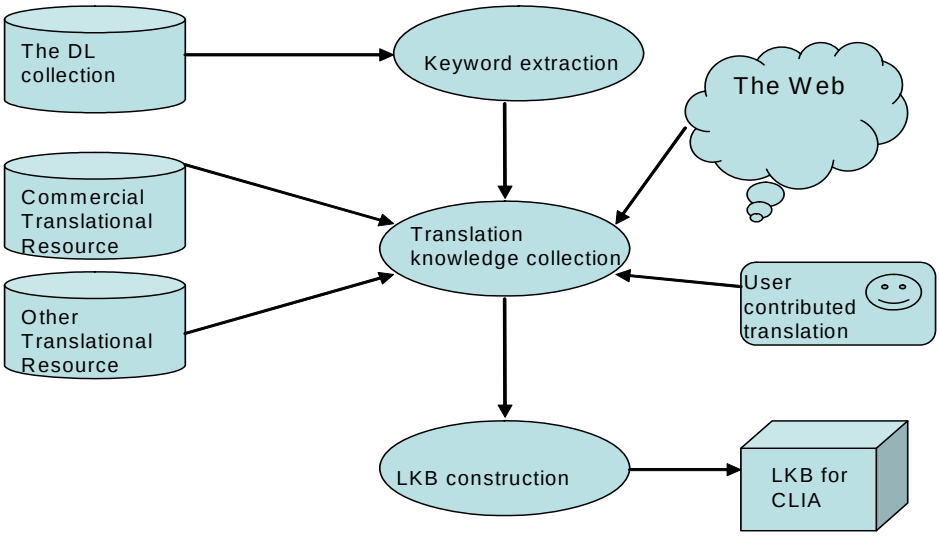


Figure 1: the LKB construction Process

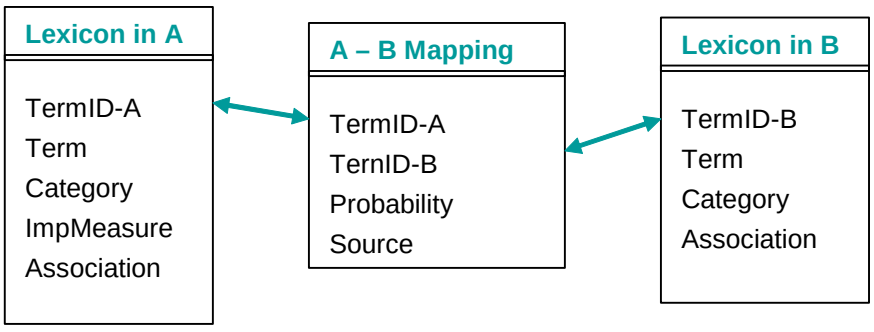


Figure 2: the Structure of the LKB

- Integrating different measures for effectiveness and efficiency. Performance and efficiency measures in system-oriented and user-oriented IR evaluation, library and

digital library service evaluation, and usability consideration should be considered in order to develop a comprehensive measure for implementing CLIA services.

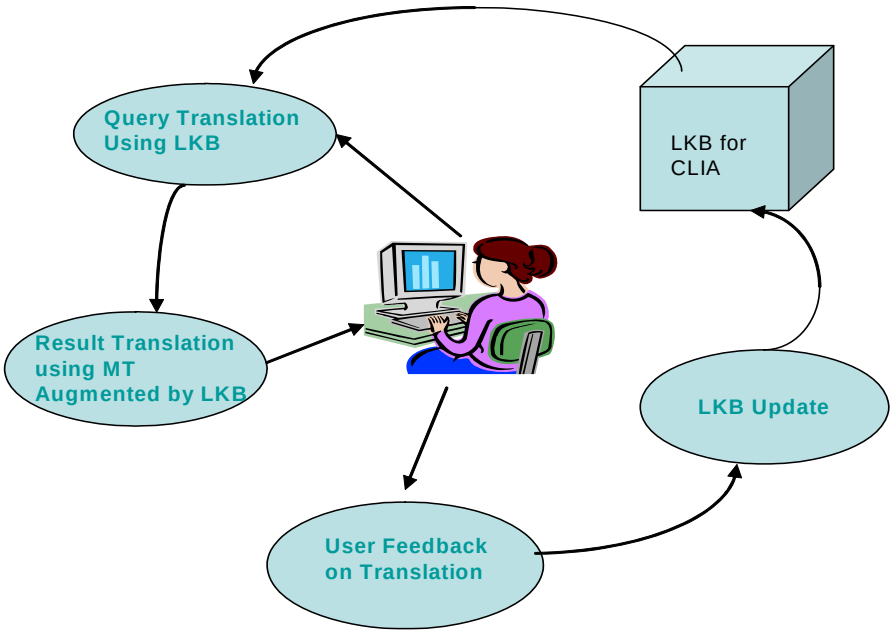


Figure 3: CLIA for DL using LKB – An Iterative & Interactive Process

3. Strategies

Collaboration is the key in order to develop the iCLIA framework. Following strategies can be considered to develop the CLIA services based on the iCLIA framework for digital libraries with limited funding:

- Collaborate with Digital library (DL) developers to work on real digital libraries. Such collaboration facilitates exploration on solutions that are appropriate for the specific digital objects while seeking funding to support CLIA as a value-added service. As digital objects are more organized than Web pages crawled by search engines, it is possible that better performance of machine translation could be achieved through the construction of the LKB for machine translation software.
- Collaborate with researchers and DL developers in other countries to increase the languages that the users can access. Many digital libraries manage precious digital assets that can be attractive to people at the other side of the earth. Collaboration with colleagues in other countries would make information resource available to use for larger population.
- Collaborate with the users. In current digital age, even monolingual digital libraries are also accessed by people who don't know the language (Sorid, 2008). Social computing has been widely used on the Internet, and it can play big role for involving users to the multilingual information access services: users may volunteer to translate

digital objects into another language. They may help to correct errors produced by machine translation systems. They can donate money to help the DL to offer the new service if they know the significance of the service, or the information needs from the other side of the earth.

Take a step-by-step approach, the DL system can be first provided with a cross-language interface, then CLIA to the metadata, and then to the full text of the digital collections.

4. Summary and Conclusion

Information systems such as digital libraries would better serve their users if language support services were integrated as part of the systems. We propose to explore an integrative CLIA framework for digital libraries. CLIA can be a value-added service to many DLs if we explore the methods, the user needs, and the evaluation of such services.

Our future research will be developing and testing the iCLIA framework to understand the application of CLIA for DLs. Within the framework, we will like to understand more about the needs and information behavior of bilingual users because bilingual users have been identified as the most possible users for CLIR systems. Also, we will collaborate with existing digital libraries to investigate effective and efficient solutions on providing multilingual information access for the digital library users.

References:

Chen, J. (2006). A lexical knowledge base approach for English-Chinese cross-language information retrieval. Journal of the American Society for Information Science and Technology, 57(2), 233-243.

Gey, F. C., Kando, N., and Peters, C. (2005). Cross-language information retrieval: the way ahead. Information Processing and Management, 41, 415-431.

Chen, J. (2007). Services provided by digital libraries: current status and future strategies. *Library and Information Service,* 51(12), 26-32.

Jones, G., Collier, N., Sakai, T., Sumita, K., and Hirakawa, H. (2001) A Framework for Cross-language Information Access: Application to English and Japanese. *Computer and the Humanities*, 35, 371-388.

Ruiz, M.E., Chen, J., Druin, A., Kando, N., Oard, D., and Peters, C. Multilingual Access to Digital Libraries. In Proceedings of the ASIS&T 2008 Annual Meeting, Columbus, OH. October 2008.

Sorid, D. (2008). Writing the Web's Future in Numerous Languages. The New York Times, December 31, 2008.

A Virtual Evaluation Track for Cross Language Link Discovery

Wei Che (Darren) Huang
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
w2.huang@student.qut.edu.au

Andrew Trotman
Department of Computer Science
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

Shlomo Geva
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT

The Wikipedia has become the most popular online source of encyclopedic information. The English Wikipedia collection, as well as some other languages collections, is extensively linked. However, as a multilingual collection the Wikipedia is only very weakly linked. There are few cross-language links or cross-dialect links (see, for example, Chinese dialects). In order to link the multilingual-Wikipedia as a single collection, automated cross language link discovery systems are needed – systems that identify anchor-texts in one language and targets in another. The evaluation of Link Discovery approaches within the English version of the Wikipedia has been examined in the INEX Link-the-Wiki track since 2007, whilst both CLEF and NTCIR emphasized the investigation and the evaluation of cross-language information retrieval. In this position paper we propose a new *virtual evaluation track*: Cross Language Link Discovery (CLLD). The track will initially examine cross language linking of Wikipedia articles. This virtual track will not be tied to any one forum; instead we hope it can be connected to each of (at least): CLEF, NTCIR, and INEX as it will cover ground currently studied by each. The aim is to establish a virtual evaluation environment supporting continuous assessment and evaluation, and a forum for the exchange of research ideas. It will be free from the difficulties of scheduling and synchronizing groups of collaborating researchers and alleviate the necessity to travel across the globe in order to share knowledge. We aim to electronically publish peer-reviewed publications arising from CLLD in a similar fashion: online, with open access, and without fixed submission deadlines.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Search Process*.

General Terms: Measurement, Performance, Experimentation

Keywords: Cross Language, Link Discovery, Information Retrieval, Evaluation

SIGIR'09, July 19–23, 2009, Boston, USA.

1. INTRODUCTION

1.1 Background

Collaborative hypertext knowledge management systems, such as the Wikipedia, offer an efficient means for creating, maintaining, and sharing information. Extensive linking between documents in these systems is essential for user navigation and assists readers to varying degrees. A reader with extensive background knowledge of a topic may be less likely to follow a link, while a less knowledgeable reader may choose to follow many links in order to expand their knowledge.

Links in the Wikipedia originate from two primary sources: the page authors' knowledge of the document collection; and automated link discovery systems such as *We Can Link It* [6] and *Wikify* [9]. The Link-the-Wiki track at INEX [22] was established as an independent evaluation forum for measuring the performance of these kinds of link discovery systems.

In 2007 the track explored document-to-document link discovery in the English Wikipedia, in 2008 the track also looked at anchor-text identification within the source document and the placement of the target point (the best-entry-point, or BEP) within the target document. This second kind of link discovery is known as anchor-to-BEP link discovery, or focused link discovery. The track also developed a standard methodology and metrics for the evaluation of link discovery systems.

The track's results show that excellent link discovery systems have been developed – that is, there are now published algorithms that can almost perfectly predict the links in a Wikipedia page. However, manual assessment revealed the highly unexpected result that many existing Wikipedia links are not relevant (a least not to the INEX assessors)! INEX now recommends manual assessment as the preferred procedure for the evaluation of link discovery systems. We note that an INEX assessor manually assessing links from the pool perfectly models a user who (after adding a new article to the Wikipedia) is navigating a list of links recommended by a link discovery system – accepting or rejecting as they go. This process lends itself to the interactive study of link discovery systems.

With the growth of the multilingual Wikipedia (and the multilingual web) there is a growing need for cross-language information retrieval including cross language interlinking of multilingual documents. Most Wikipedia pages are written in English and we, unsurprisingly but anecdotally, observe users whose first language is not English searching the English Wikipedia. Their need is two-fold: the Wikipedia documents to be translated into their first language; and links between

documents to reflect their language choices. Translation is already happening and some cross-language links already exist, however these problems are our research motivation. We are trying to:

- Identify Wikipedia documents that are all on the same topic, irrespective of language, and
- Identify hypertext links between documents, irrespective of language, so that a user can choose a target document based on a language preference.

1.2 Motivation

Many Internet users are multi-lingual. To satisfy their information needs the search engine should return documents in the different languages they read. Doing so is more thorough than returning results in just one language. As examples, the *Early history of the United States of America* can be found in the Chinese Wikipedia but the English Wikipedia has a much richer document on the topic; information about *Chinese Dynasties* may be found in several documents in the Chinese language Wikipedia, and indeed in several distinct Chinese language version of the Wikipedia. In both examples, a link between these different language versions will help the multi-lingual reader. In both examples focused cross-lingual anchor-to-BEP links would result in a more comprehensive interlinked knowledge base, especially if the links the multilingual reader sees are based on a personal language profile. Envisage a document being interlinked to any number of languages, but users only seeing links to languages that are defined in their personal profile.

Anchor-to-BEP linking is a feature of HTML that is rarely exploited in links – despite its existence since the beginnings of the web. Very few links in the Wikipedia actually take the user from the point of reference (the anchor) to the target location within another document (the BEP). Such interlinking is common within a single document and is used in navigation, but is rarely utilised when linking between documents. Such *focused* interlinking is particularly desirable when documents are large or when browsing on small mobile devices. For instance, in the article *South Eastern Main Line*, an orphaned anchor, *Folkestone Harbour*, is colored in red. It is a place-holder for a link to an article that does not yet exist. However, the article *Folkestone* does have a section titled *Folkestone Harbour*. This prospective anchor could be linked to this section until an article on Folkestone Harbour is created.

Extending the INEX Link-the-Wiki track to cross language linking will help turn the Wikipedia into a multi-lingual knowledge network. The section 地理 (English: *Geography*) in the article 英国 (English: *England*) has two anchor texts, 多佛港 (English: *Dover Harbour*) and 英吉利海峡隧道 (English: *Channel Tunnel*; French: *Le tunnel sous la Manche*). There is no link for 多佛港 (English: *Dover Harbour*) in the Chinese Wikipedia, but an article on the *Port of Dover* is linked from the redirect of the *Dover Harbour* page in the English Wikipedia. Information (images and geography) about Dover Harbour can also be found in the English article on *Dover*. The article, 英吉利海峡隧道, does not express much information about the channel tunnel, certainly not as much as the English *Channel Tunnel* page. These two examples show the need for cross-language links within the Wikipedia.

To the best of our knowledge, current link discovery systems such as *Wikify*[9] focus on monolingual Wikipedia and have not been extended to support multilingual link discovery. Cross-language tracks conducted in NTCIR and CLEF explore Information Retrieval and Question-Answering but not link discovery. Link Discovery is different from Information Retrieval although it does rely on similar technology: for link discovery a match of semantic context between the point of reference (the context of the anchor) and the target text (the BEP context) is essential. Unlike query based information retrieval, in link discovery the context of the anchor is always explicit since the anchor is always embedded in surrounding text, and similarly the context of the target location is highly focused and specific. In information retrieval the query is known but the context unknown, in link discovery it is necessary to identify both the query (the anchor-text) and the results list (the target document and BEP) and embedding contexts are available.

Link discovery provides a rich context in which NLP based approaches may well prove much more useful than they had been in the query based information retrieval. Furthermore, cross-language link discovery involves a set of technologies, including IR, NLP, semantic and similarity matching techniques, character encoding technologies, machine readable corpora and dictionaries, machine translation, focused and passage retrieval, and multiple links per anchor discovery. Cross Language Link Discovery (CLLD) demands the tight integration of techniques currently under examination at INEX, CLEF and NTCIR.

Herein we formally propose the CLLD track. This track will be run as a single collaborative web-based forum. Participants will be drawn from the existing forums, but be part of none (or all), it will be an online *virtual evaluation forum*. All collections, topics, submission and result analysis will be maintained via a remote repository. By using only public domain data (such as the Wikipedia and open source software) we can simplify participation and the sharing of resources. The community of participants will provide software tools and assessments; as well as a peer-reviewed online publication for approaches and results. The forum will not be tied to any particular timeline or venue but will be run as a continuous evaluation track – without a dedicated annual event (although there is no reason not to hold such meetings, perhaps as surrogate to larger events). This proposal represents a dramatic philosophical change from the traditional TREC paradigm.

2. RELATED WORK

As suggested by Wilkinson & Smeaton [1], navigation between linked documents is a great deal more than simply navigating multiple results of a single search query, linking between digital resources is becoming an ever more important way to find information. Through hypertext navigation, users can easily understand context and realize the relationships in related information. However, since digital resources are distributed it has become difficult for users to maintain the quality and the consistency of links. Automatic techniques to detect the semantic structure (e.g. hierarchy) of the document collection, and the relatedness and relationships of digital objects have been studied and developed [2]. Early works, in the 1990s, determined whether and when to insert links between documents by computing document similarity. Approaches such as term repetition, lexical chains, keyword weighting and so on were used to calculate the similarity between documents [3, 4, 5]. These approaches were

focused on the document-to-document linking scenario, rather than identifying which parts of which documents were related.

Jenkins [6] developed a link suggestion tool, *Can We Link It*. This tool extracts a number of anchors which have not been discovered in the current article and that might be linked to other Wikipedia documents**.** The user can accept, reject, or click *"don't know"* to leave a link as undecided. Using this tool the user can add new anchors and corresponding links back to a Wikipedia article.

A collaborative knowledge management system, called *PlanetMath*, based on the *Noosphere* system has been developed for mathematics [7]. It is encyclopedic, (like the Wikipedia), but mainly used for the sharing of mathematical knowledge. Since the content is considered to be a semantic network, entries should be cross-referenced (linked). An automatic linking system provided by Noosphere employs the concept of conceptual dependency to identify each entry for linking. Based on the Noosphere system, *NNexus* (Noosphere Networked Entry eXtension and Unification System) was developed to automate the process of the automatic linking procedure [8]. This was the first automatic linking system which eliminates the linking efforts required by page authors.

The *Wikify* [9] system which integrates technologies of automatic keyword extraction and word sense disambiguation can identify the important concepts in a document and link these concepts to corresponding documents in the Wikipedia. Mihalcea and Csomai stated that many of applications such as the annotation of semantic webs and academic systems can benefit from this kind of system.

Since the inception of TREC in 1992 interest in IR evaluation has increased rapidly and today there are numerous active and popular evaluation forums. It is now possible to evaluate a diverse range of information retrieval methods including: ad-hoc retrieval, passage retrieval, XML retrieval, multimedia retrieval, question answering, cross language retrieval, link discovery, learning to rank, and so on.

The CLIR (Cross-Language Information Retrieval) track was first introduced to TREC in 2000. It offered document collections in English, French and German and queries in English, French, German, Spanish and Dutch. Three fundamental resources, machine translation, machine readable dictionaries and corpus-based resources, have been used. There are three common approaches to match queries with the resource documents [10]: machine translation technology using dictionaries and statistical information or example-based translation; machine readable bilingual dictionaries; and relying on corpus resources to train retrieval mechanism by using either Latent Semantic Indexing (LSI), Generalized Vector Space Model or Similarity Thesauri for translating queries. As a performance baseline corresponding language queries were also submitted against the same language collections.

The cross-language track investigated the retrieval of relevant documents that pertain to a particular topic or query regardless of the language in which both the topic and documents were written. The last TREC CLIR track was run in 2002, whoever ongoing effort can be found in both NTCIR and CLEF [11].

NTCIR started in late 1997 and is focused on Japanese and other East Asian languages [12]. The workshop runs on an 18-month cycle. The aim is to build an infrastructure for large-scale experimental evaluation of Information Access (IA) research. IA in the workshop has been indicated as the process of searching, browsing and looking for relevant information, and utilizing the information. Technologies, like Information Retrieval (IR), Cross-Language Information Retrieval (CLIR), Question-Answering (QA), text summarization and text mining, are considered part of the IA family. In 2008 (NTCIR-7), there were 6 tasks: Complex Cross-Lingual Question Answering (CCLQA), Information Retrieval for Question Answering (IR4QA), Multilingual Opinion Analysis Task (MOAT), Patent Mining (PAT MN), Patent Translation (PAT MT) and Multimodal Summarization of Trend (MuST).

The goal is to develop a module-based infrastructure for evaluation integrating IR and QA technologies to propose answers in a suitable format for given questions in any language. It intends to model significant work from every participant and build a set of APIs (or modules) to facilitate the development of cross-language (QA) systems. A platform, called EPAN (Evaluation Platform for ACLIA and NTCIR), was adopted by NTCIR to perform the collaborative evaluation [13]. Through module-by-module evaluation it is possible to identify problems in parts of participants' otherwise complicated CLIR-QA systems – something not possible in end-to-end evaluation. For example, many CLIR-QA systems failed to retrieve relevant documents when named entities in queries did not appear in ordinary translation dictionaries. The module-based approach also makes it possible for participants to collaborate by working on different modules.

The annual CLEF forum aims to create a research forum in the field of multilingual system development [14]. The experiments range from monolingual text retrieval to multilingual multimedia retrieval. The collection and available languages vary depending on different tasks. They include 3 million News articles in 13 languages, a social science database in English and German, the Cambridge Sociological Abstracts, and the Russian ISISS collection, 3.5 million web pages in multi-languages, and a photograph database with captions in different languages [15]. Various sets of topics in different languages are available for respective tasks.

The DIRECT system used by CLEF is manages data (collection, topics, and metrics), building statistics (analysis, plots and results) and provides different entries for various roles [16, 17]. Of particular note is the dynamic user interface through which participants can to interact with their-own and others' experimental data and results.

3. MULTILINGUAL WIKIPEDIAS

The Wikipedia is a multilingual online encyclopedia that offers a free and flexible platform for developing a collaborative knowledge repository [18]. Currently, it has entries written in more than 200 different languages. Overell [19] shows that the geographic coverage of the Wikipedia very much depends on the language version – places in the UK are best covered by the English language version of the Wikipedia while places in Spain are best covered by the Spanish language version. There are more than 2,874,919 articles (May 2009) in the English Wikipedia which is the largest language version in the Collection. By the end of 2008, no fewer than 13 languages have more than 200,000 articles. As can be seen in Table 1, both European and Asian

language versions have reached a substantially size – the collection is already useful for multilingual research[20].

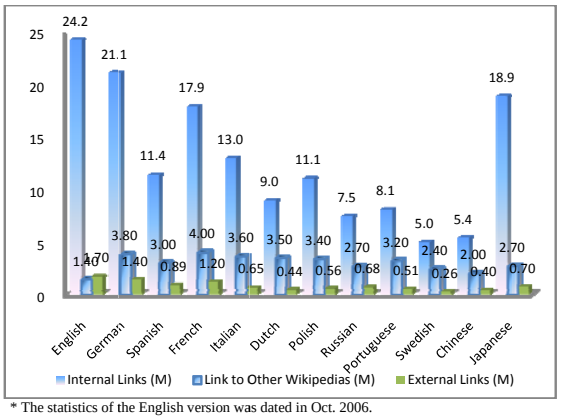
Table 1: Language subsets of the Wikipedia at end 2008

	Articles (K)	Database (GB)	Bytes per Article	Over 2KB
English*	1100	3.2	3092	363 (33%)
German	858	3.5	3489	420 (49%)
French	746	2.7	2995	269 (36%)
Polish	567	1.5	1988	130 (23%)
Japanese	555	2.6	1890	128 (23%)
Italian	533	1.9	2914	192 (36%)
Dutch	507	1.3	2021	147 (29%)
Portuguese	448	1.1	1866	90 (20%)
Spanish	430	1.8	3492	189 (44%)
Russian	347	1.9	2822	125 (36%)
Swedish	301	0.609	1535	54 (18%)
Chinese	209	0.717	1451	36 (17%)
Norwegian	203	0.475	1781	43 (21%)

* WikiMedia state the English version as at Oct. 2006!

Most documents have a rich set of same-language links but a much lower number cross-language links. Figure 1 shows that by October 2006 (WikipMedia have not provided stats for English since that date) only 1.4 million English anchor texts have been linked to entries in other languages while the amount of same-language links in the English Wikipedia had reached 24.2 million.

Around a quarter of anchor texts in the Chinese Wikipedia were linked to other language collections in the Wikipedia. There are several different Chinese dialects available as different collections and many of English terms are linked to the English Wikipedia. Around 50% of Wikipedia documents in German version are linked to the English Wikipedia whilst only 14% of English documents have been linked to the German version [21]. Other versions in Figure 1 have less than 20 percent of their links pointing to other languages. Linking the Wikipedia entries across different languages is still very limited. Efficient and accurate cross language retrieval is yet to be demonstrated and evaluated.



* The statistics of the English version was dated in Oct. 2006.

Figure 1: Different linking types in different language versions of the Wikipedia at the end of 2008

In order to achieve a comprehensive cross-language knowledge network, cross language link discovery is essential. The document collection is large (there are many millions of documents) and broad (there are many languages covered). Although only one link per anchor is typically displayed by existing HTML based web browsers, there is no inherent restriction for this limit in the HTML standard. Anchor text can be linked not only to multiple targets, but also in different languages, and the extension of browsers to support this functionality is long overdue.

4. INEX *LINK-the-WIKI* TRACK

The INEX Link-the-Wiki track offers a standard forum for the evaluation of link discovery in both document-to-document and anchor-to-BEP linking. The task is to discover a set of anchor texts, based on the context of the topic, which can semantically be linked to best entry points in other documents. Besides outgoing links, candidate incoming links from other articles into the topic document are also required.

The INEX 2008 English Wikipedia collection consisting of 659,388 documents was used as the corpus for the experiments since Wikipedia is composed of millions of interlinked articles in numerous languages and have been proved as a valuable collection for IR research. 50 documents nominated by participants were used in the anchor-to-BEP task. For the document-level link discovery, 6,600 documents were randomly selected but pre-filtered (for suitability) by size and the number of links. For 2009 this collection has been updated and now consists of over 2.6 million articles, spanning over 60GB of text (without images).

The documents were converted into topics by removing all links in the documents and removing all links from the remainder of the collection to those documents, they were said to have been orphaned from the collection. The original documents (with the links) were said to be pre-orphans.

Links within the collection to and from the pre-orphans were extracted and used as the ground truth to which runs were compared. This is the automatic evaluation method. Using standard methodology, the topics were sent to participating groups. Participant's link discovery systems were required to return a *ranked* list of at most 50 outgoing text anchors (each of which targeted up-to 5 documents). The results were exhaustively pooled and the ground truth of existing links in the Wikipedia was added to the pools. Pools were manually assessed to completion, by the groups that nominated the topics. This formed the manual evaluation set. Assessment against the ground-truth result set generated a score for the performance of a submission relative to the Wikipedia. Additional evaluation was performed against the manually assessed links (including the manually assessed existing Wikipedia links). The outcome showed that the Wikipedia ground-truth does not agree with user expectation. The manual assessment process is necessary in order to produce a reliable test base for evaluation.

5. PROPOSED CLLD METHODOLOGY

The cross language link discovery track will work in a similar manner to the INEX track. Participating groups will be asked to submit a list of languages they can read (from those covered by the Wikipedia). This also indicates the languages these participants can submit runs for and can assess in. A subset of topics from the different language Wikipedia collections will be chosen and distributed to participants. To support the creation of topics, a selection tool will be developed to help choose and to orphan documents in indicated languages. Participants will be able to choose any combination of languages, for example German documents linking to English or Dutch documents, or Chinese documents linking to Japanese documents.

Programs will be provided that allow participants to view their runs. These programs will show proposed linked documents with the anchors and their respective target document best-entry-points, as would be seen by an assessors (and ultimately end-users of their system).

The challenge for the organizers is to obtain a critical mass of participants and assessors to facilitate robust and reliable manual evaluation in multiple languages. The track must, therefore, be a close and extensive collaboration between NTCIR, CLEF, INEX, and other evaluation forums.

5.1 Tasks Specification

Initially two linking tasks will be formalized:

- MULTILINGUAL topical linking:
  This is a form of document clustering – the aim is to identify (regardless of language) all the documents in all languages that are *on the same topic*. The Wikipedia currently shows these links in a box on the left hand side of the page.
- BILINGUAL anchor linking:
  It is exemplified by the Chinese article 诺森伯兰郡, having a link from the anchor 国会选区 to the English article *List of United Kingdom Parliament constituencies*. The link discovery system must identify the anchor text in one language version of the Wikipedia and the destination article within any other language version of the Wikipedia.

In the case of MULTILINGUAL topical linking, the participants are encouraged to discover as many documents as they can.

In the case of BILINGUAL anchor linking, at most 50 anchors may be identified in a orphaned document and up to 5 BEPs may be linked to one target language (e.g. English to German). Initially only outgoing links will be examined since incoming links from a single language may not make sense.

5.2 Test Collections and Submission

The set of multilingual Wikipedia collections will be used as the corpus for cross language link discovery. The size and the number of documents is listed in Table 1. Nominated topics will be collected and the ground-truth extracted from the collection.

Participants will be encouraged to share in the development of appropriate procedures for topic selection, multilingual topic discovery, ground truth link extraction, and the assessment method.

The submission format may be derived from the format currently used by INEX. The existing INEX tools will be ported to support CLLD.

5.3 Evaluation

It is essential to define a standard methodology and procedure to assess link quality and to quantitatively evaluate different approaches.

5.3.1 Static Evaluation

When Trotman & Geva [24] introduced the Link-the-Wiki track at INEX they also noted that the evaluation required no human assessment. The same is true with cross-language link discovery.

Topics in the INEX Link-the-Wiki track are chosen directly from the document collection. All links in those documents are removed (the document is orphaned). The task is to identify links for the orphan (both from and to the collection). Performance is measured relative to the pre-orphan (the document before the links were removed).

For MULTILINGUAL linking the links on the left hand side of the Wikipedia page could be used as the ground truth. The performance could be measured relative to the alternate language versions of the page already known to exist.

BILINGUAL anchor linking from one document to another could also be automatically evaluated. Links from the pre-orphan to a destination page in an alternate language would be used as the ground truth – but there are unlikely to be many such links.

A same-language link from a pre-orphan to a target provides circumstantial evidence that should the target exist in multiple languages then the alternate language versions are relevant. This is essentially a triangulation: $A \xrightarrow{t} B \xrightarrow{l} C \Rightarrow A \xrightarrow{tl} C$ where *A*, *B*, and *C* are documents; and *t* designates a topical link, *l* a cross language link, and *tl* a topical cross language link. By extension, if $A \xrightarrow{t} B \xrightarrow{T} C \Rightarrow A \xrightarrow{tT} C$, where $B \xrightarrow{T} C$ designated two documents that are not linked but have the same title.

Static assessment requires no human interaction. A web site with orphan sets chosen using some criteria (perhaps randomly), with the assessment sets (extracted from the pre-orphans), and that will evaluate a run will be built and provided. Such an evaluation methodology raises the possibility of running the track continually and without any deadlines.

5.3.2 Continual Evaluation

Huang *et al*. [22] question automatic evaluation. Their investigation suggests that many of the links in the Wikipedia are not topical, but are trivial (such as dates), and that users do not find them useful. Manual assessment is, consequently, necessary. This raises new challenges for cross language link discovery because finding assessors fluent in multiple languages is difficult – especially for a track with a relatively small number of participants but in a large number of languages (the Wikipedia has over 200 languages).

We propose a novel form of evaluation called *continual evaluation* in which participants can download topics and submit runs at any time; and in which contribution to manual assessment is an on-going concern. The document collection will, initially, be static. Topics will either be chosen at random from the collection, or nominated by participants. For any given run a participant will download a selection of topics and submit a run. The evaluation will be based on metrics that consider the un-assessed document problem (such as a variant on rank-biased precision [23]), and comparative analysis will be relative to an incomplete, but growing, assessment set.

To collect assessments two methods are proposed: first, in order to submit a run the participant will be required to assess some anchor-target pairs in languages familiar to them; second, we will run an assessment Game With A Purpose (GWAP). Kazai et al. used a GWAP for the INEX Book track; Von Ahn & Dabbish [25] discuss GWAPS in other contexts (including the Google Image Labeler). Regardless of the method of assessment collection, we are trying to validate the minimum number of links necessary to disambiguate the relative rank order of the runs (within some known error).

6. PUBLICATION

Both automatic and manual assessment of cross language link discovery can be done on a continual rolling basis; there is no need for topic submission deadlines, run deadlines, assessment deadlines; or paper publication deadlines.

At INEX the time difference between run-submission and the workshop paper submission date is long (6 July – 23 Nov). With automatic assessment it is possible to achieve a result, write, and then publish a paper with a short turn around. As part of the virtual track we propose an open-access virtual workshop workbook to which registered participants can immediately submit their papers for peer-review and publication.

7. CONCLUSIONS AND OUTLOOK

As far as we are aware, the cross-language link discovery track is the first to offer extensive reusable independent evaluation resources. In this paper we introduce this new evaluation task.

A fully automated procedure for anchor-to-document link analysis, using the existing Wikipedia linking network is described. The procedure was used at INEX 2007 and allowed us to create a fast evaluation procedure with a turnaround time of days and not months because it had no manual assessment. The procedure allows for a very large number of documents to be used in experiments. This overcomes the assessment bottleneck which is encountered in most other tasks in collaborative evaluation forums such as INEX and TREC. We further proposed to extend the task to Cross Language Link Discovery, and propose the concept of automatic evaluation. We describe the requirement for evaluating such a task.

These activities may not be held in a fixed place but can be done by gathering participants from INEX, CLEF and NTCIR through a virtual web-based system. The CLLD track will be dedicate to supporting efficient methods and tools for CLLD evaluation. The collections, submission and result data will be well managed for further analysis and experiments. Participants from different nations are expected to work collaboratively to achieve the development of multilingual link discovery systems.

Baseline automatic evaluation methods seen at INEX do not require human intervention as the assessments are extracted directly from the collection and performance is measured relative to these. The new track can, therefore, bootstrap and run online with continuous evaluation, free from the problems of scheduling groups of collaborating researchers. Overtime manual assessments will be collected and improve the available resources. We also propose to publish the results of the track in a similar fashion to the CLLD track itself – online, with open access, and with quality control.

8. REFERENCES

[1] Wilkinson, R. and Smeaton, A. F., *Automatic Link Generation*, ACM Computing Surveys, 31(4), December 1999.

[2] Green, S. J., *Building Hypertext Links By Computing Semantic Similarity*, IEEE Transactions on Knowledge and Data Engineering, September/October 1999, 11(5), pp. 713-730.

[3] Allan, J., *Building Hypertext using Information Retrieval*, Information Processing and Management, 33(2) pp. 145-159.

[4] Green, S. J. (1998) *Automated Link Generation: Can We Do Better than Term Repetition?*, In Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 75-84.

[5] Zeng, J. and Bloniarz, O. A. (2004) *From Keywords to Links: an Automatic Approach*, In Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), 5-7 April 2004, pp. 283-286.

[6] Jenkins, N., *Can We Link It*, http://en.wikipedia.org /wiki/User:Nickj/Can_We_ Link_It.

[7] Krowne, A, *An Architecture for Collaborative Math and Science Digital Libraries*, Thesis for Master of Science Virginia Polytechnic Institute and State University, 19 July 2003.

[8] Gardner, J., Krowne, A. and Xiong, L., *NNexus: Towards an Automatic Linker for a Massively-Distributed Collaborative Corpus*, In Proceedings of the International Conference on Collaborative Computing, 17-20 November 2006, pp. 1-3.

[9] Mihalcea, R. and A. Csomai, *Wikify!: linking documents to encyclopedic knowledge*, CIKM 2007, pp. 233-242.

[10] Schäuble, P. and Sheridan, P., *Cross-Language Information Retrieval (CLIR) Track Overview*, In Proceedings of the Sixth Text Retrieval Conference, pp. 31-44.

[11] Agosti, M., Di Nunzio, G. M., Ferro, N., Harman, D. and Peters, C., *The Future of Large-Scale Evaluation Campaigns for Information Retrieval in Europe*, LNCS 4675, 2007, pp. 509-512.

[12] Kando, N., *Overview of the Seventh NTCIR Workshop*, In Proceedings of NTCIR-7 Workshop Meeting, 16-19 December 2008, Tokyo Japan, pp. 1-9.

[13] Mitanura, T., Nyberg, E. Shima, H., Kato, T., Mori, T., Lin, C.Y., Song, R., Lin, C. J., Sakai, T., Ji, D. and Kando, N., *Overview of the NTCIR-7 ACLIA Taska: Advanced Cross-Language Information Access*, In Proceedings of NTCIR-7 Workshop Meeting, 16-19 December 2008, Tokyo Japan, pp. 11-25.

[14] Ferro, N. and Peters, C., From *CLEF to TrebleCLEF: the Evolution of the Cross-Language Evaluation Forum*, In Proceedings of NTCIR-7 Workshop Meeting, 16-19 December 2008, Tokyo Japan, pp. 577-593.

[15] Di Nunzio, G. M., Ferro, N., Mandl, T. and Peters, C., *CLEF 2007: Ad Hoc Track Overview*, CLEF 2007, LNCS 5152, pp. 13-32.

[16] Di Nunzio, G. M. and Ferro, N., *DIRECT: A System for Evaluating Information Access Components of Digital Libraries*, ECDL 2005, LNCS 3652, pp. 483-484.

[17] Dussin, M. and Ferro, N., *Design of the User Interface of a Scientific Digital Library System for Larger-Scale Evaluation Campaigns*, Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008), pp. 105-113.

[18] Wikipedia, the free encyclopedia, 2009. http://wikipedia.org/.

[19] Overell, S.E., *Geographic Information Retrieval: Classification, Disambiguation and Modelling*, in *Department of Computing*. 2009, Imperial College London: London. pp. 175.

[20] Wikipedia Statistics, Wikipedia, the free encyclopedia, 2009, http://stats.wikimedia.org/EN/Sitemap.htm.

[21] Sorg, P. and Cimiano, P., *Enriching the Cross-lingual Link Structure of Wikipedia – A Classification-Based Approach* -, In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence, June 2008.

[22] Huang, W.C., A. Trotman, and S. Geva, *The Importance of Manual Assessment in Link Discovery*, in *SIGIR 2009*. 2009, ACM Press: Boston, USA.

[23] Moffat, A. and J. Zobel, *Rank-biased precision for measurement of retrieval effectiveness.* ACM Trans. Inf. Syst., 2008. 27(1):1-27.

[24] Trotman, A. and S. Geva. *Passage Retrieval and other XML-Retrieval Tasks*. In SIGIR 2006 Workshop on XML Element Retrieval Methodology. 2006. Seattle, Washington, USA.pp. 43-50.

[25] von Ahn, L. and L. Dabbish, *Designing games with a purpose*. Commun. ACM, 2008. **51**(8):58-67.

Urdu is not Hindi for Information Access

Kashif Riaz
University of Minnesota
4-192 EE/CS Building
200 Union Street SE
Minneapolis, MN 55455
riaz@cs.umn.edu

ABSTRACT
Urdu and Hindi have a complex relationship. Urdu is written in the Arabic script, and Hindi is written in Devanagri script. The relationship between these two languages is deep-rooted based on geo-political and historical events. Urdu and Hindi are considered very similar languages, and sometimes linguists refer to them as Hindi-Urdu. Urdu is considered a "scarce resource" language but Hindi has a vibrant toolset to do research in Information Retrieval. In this position paper I contend that language resources and enabling technologies for Information Access cannot be used interchangeably between these two languages. More specifically, Hindi cannot be used as bridge language to do research in Information Retrieval in Urdu. I argue this assertion using deep analysis of the language through socio-linguistics and quantitative analysis using Zip's Law. The contrast and comparison are done using script and vocabulary differences between the two languages.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: Linguistic Processing, Dictionaries, Indexing methods.

General Terms
Algorithms, Standardization, Languages

Keywords
Urdu, Hindi, Script, Orthography, Language Resources

1. INTRODUCTION
Urdu and Hindi share a complex relationship. Together they boast one of the largest populace who understands them and call either one of them as their national language. They are the languages of South Asia—Urdu is the national language of Pakistan, and Hindi is the official language of India. India does not have a national language because of the number of regional language-sensitive provinces. Urdu is one of the official languages of India. Urdu is written in the Arabic script, and Hindi is written in Devanagri script. While doing research on Urdu named entity recognition, I wanted to use some Hindi language resources like gazetteers and online dictionaries since they are not available for Urdu. I realized in early research stages that Hindi and Urdu were behaving as separate languages. Moreover, I needed to learn Devanagri script to proceed further. In this position paper I argue that Hindi and Urdu, although grammatically very similar, cannot be treated as the same language for doing research in computational linguistics and information retrieval—at least with the current set of tools available for both languages. Hindi has quite a vibrant set of enabling technologies for Information Access whereas research in Urdu is still in its infancy. Some of the examples of these enabling technologies are: online dictionaries, Wordnet, stemmers, stop word lists, gazetteers for named entity recognition, part of speech taggers, baseline for evaluation, etc. Some of these tools exist in Hindi but not in Urdu; if they exist in Urdu, they are rudimentary. Some tools for Urdu are available through CRULP [4] and [1, 3, 9, 11, 12]. Section 2 gives a brief overview of Urdu. Section 3 analyzes the differences between Urdu and Hindi and motivates why a position needs to be taken for doing Information Retrieval. Section 4 gives three examples of quantitative analysis between Hindi and Urdu to reinforce the position. Section 5 concludes the position paper with remarks.

SIGIR'09, July 19-23, 2009, Boston, MN, USA.

2. Urdu
In this section I will briefly introduce some right-to-left languages and a few characteristics of Urdu. Urdu is the national language of Pakistan and one of the major languages of India. It is estimated that there are about 300 million speakers of Urdu. Most Urdu speakers live in Pakistan, India, the UAE, the U.K., and the USA. Urdu is considered the *lingua franca* of business in Pakistan and amongst the South Asian community in the U.K. [3]. Personally, when travelling to the U.K., I rarely speak English in big cities like London or Glasgow while out and about.

Urdu has a property of accepting lexical features and vocabulary from other languages, most notably, English. This is called code-switching in linguistics. It is not uncommon to see a right-to-left text flow interrupted by a word written in English left to right and then a continuation of the flow right to left. For example, وہ میرا laptop ہے [That is my laptop]. In the above example, Microsoft Word did not support English embedding within the Urdu sentence, and displayed it improperly, but while electronically processing, the tokenization will be done correctly [10].

Recently there has been quite a bit of interest in right-to-left language processing in the IR community—specifically in the intelligence community and other organizations working for the government in the United States. Most of the interest has been focused toward Arabic, Urdu, Persian (Farsi), and Dari. Arabic is a Semitic language, and the other languages belong to the Proto Indo Iranian branch of languages. Arabic and these other languages only share script and some vocabulary. Therefore, the language-specific task done for Arabic is not applicable to these languages. For example, stemming algorithms generated for Arabic will not work for a language like Urdu.

Urdu, among all above languages mentioned, has unique case in that it shares its grammar with Hindi. The difference is some vocabulary and writing style. Hindi is written in Devanagri script. Because of the grammatical similarity, Hindi and Urdu are

considered one language with two different scripts by many linguists. I will argue in later sections that there is a growing number of dissenters among South Asian language researchers. Urdu is quite a complex language because Urdu's grammar and morphology is a combination of many languages: Sanskrit, Arabic, Farsi, English, and Turkish to name a few. This aspect of Urdu becomes quite challenging while building Information Access systems [9]. Because of its rich morphology and word-borrowing characteristics, Urdu is widely considered the language of the poets. From the Mughal courts of the 1700s to freedom writings of the 1900s, it was a prerequisite to learn Urdu in order to be considered a reputable poet or intellectual.

3. Analysis of Hindi and Urdu

3.1 Background

While doing research on Urdu Information Access, a claim can be made that Urdu and Hindi are the same language in two different scripts. According to this theory, any computational model or algorithm that works for Hindi should also work for Urdu. Some examples are evaluation methodologies that are designed for Hindi, gazetteers for name entity recognition, and Wordnet for Information Access. The following section describes in detail that the *one language two scripts* theory for Urdu and Hindi is invalid in all circumstances and specifically for computational processing.

Although a lot of research has been done about the origins of Urdu and Hindi, no research study exists that compares and contrast Urdu and Hindi in a scholarly fashion [13]. For quite some time, Urdu and Hindi were treated as the same language and, indeed, they are very similar. For geo-political reasons, the languages can be classified as two languages. The geo-political reasons are of no concern to this study, but they play an important role in why these two languages are currently diverging. Only linguistic and pragmatic reasons should be considered while studying the nature of Hindi and Urdu and their impact on computational processing. I will also exclude the socio-linguistic aspects ***in terms of religion*** (i.e., Hindi for Hindus and Urdu for Muslims in India) because this leads the discussion elsewhere.

With my linguistic training, I used to think that Urdu and Hindi were the same language and differed in vocabulary only. Hindi vocabulary emerges from Sanskrit, and Urdu vocabulary borrows from many other languages, but the majority of the borrowing comes from Persian and Arabic. Some rudimentary experiments for computationally recognizing names show that Hindi and Urdu behave as two different languages. For example, lexical cues of recognition of locations are different. For example, *Dar-al-Khilafah* (Urdu) and *Rajdihani (Hindi)* are both used for the capitol of a city or a country. Therefore, more research is warranted to understand the relationship between these two languages to understand if the computational models based on one language can be used in some capacity for the other language. For this study, my reasoning is based on the analysis of some of the revered scholars of the Urdu and Hindi languages, like Ralph Russell, David Matthews, Robert King, and Intizar Hussain.

3.2 Hindustani is not the Predecessor of Hindi and Urdu

Some researchers claim that Urdu and Hindi emerged as the offspring of the earlier language called Hindustani. Although there is some reference to a language called Hindustani in the later 1800s and early 20th century, a deeper analysis shows that no such language existed, and if it did, it was Urdu. Urdu was chosen as the official language of India by the British—that changed only after the Mutiny of 1857 [6]. The use of the word "Hindustani" emerged from the leaders of the Congress Party, primarily Gandhi, and then Nehru, who wanted a common language for the united India. From the early 1900s, Gandhi relentlessly pursued the theory that the people of northern India spoke neither Persian-based Urdu nor Sanskrit-based Hindi; instead they spoke Hindustani. The reason for this argument by the party leaders was that there needed to be one language for the united India. Gandhi, and later Nehru, never talked about which script would be chosen because of the intense emotions attached to the issue [6]. This political desire did not culminate as planned because most of the followers of the Congress Party wanted the Devanagri script to be the official script for Hindustani, an idea which was rejected by the mostly Urdu-speaking populace in northern India and Hyderabad. The use of the term *Hindustani* after the partition of India and Pakistan was used to describe the political tension around language choice during the British Raj.

3.3 Divergent Trend

Hindi and Urdu have two very clear differences: script and vocabulary. I will discuss the vocabulary differences in this section. Recent research in Urdu and Hindi studies is consistently arguing that Hindi and Urdu are two different languages and that they continue to diverge as time goes on. The most notable example of continual diverging is the ultra *sanskritizing*[1] of Hindi so much so that an Urdu speaker does not understand a Hindi news broadcast. This does not mean that Urdu and Hindi speakers don't understand each other; they do at an everyday level. Also, both Hindi and Urdu speakers who live together in Uttar Pradesh, Andhra Pradesh, and other large cities understand both languages effortlessly. But a person from an outside region who travels to these areas and knows only one language does not understand the other language [6][8][13]. Matthews [8] and Martynyuk [7] make a similar comparison with Russian, Ukrainian, and Serbo-Croatian languages in terms of their similarities but still being different languages. King [6] claims that the relationship between Urdu and Hindi is much more complex than Cyrillic orthography-based Serbian and romanized Croatian collectively known as Serbo-Croatian. Martynyuk [7] agrees with Matthews [8] in general principle but disagrees that Russian and Ukrainian are similar. Instead he cites many examples of different language families like Polish and Ukrainian who have more common vocabulary. The difference between the vocabulary of Russian and Ukrainian is 38 percent and the difference between Polish and Ukrainian is 30 percent. Many more examples can be seen in [7]. Similarly, Matthews [8] claims that the situation of Hindi and Urdu is similar to Russian and Ukrainian where the two languages are perceived to be more similar than they are because of their history and native speakers living together. Given the discussion above, common vocabulary claim cannot be used to claim that two languages are same.

The relationship between Hindi and Urdu is very complex. While analyzing the differences at a high level, they can be treated as the same language and play a pivotal role in establishing a link between South Asian communities around the world. A glowing example of this phenomenon is the Indian cinema where the line

[1] Sanskritization is defined by anthropologist as spread of Vedantic and Brahmanical culture.

between Hindi and Urdu gets diluted. Although Hindi movies are popular in Urdu-speaking Pakistan and Pakistani TV shows are popular in India, there is steady and noticeable shift in Indian movies towards *sanskritizing* of Indian cinema. Remarkably, the Indian movies produced from the 1950s to the 1980s are undoubtedly Urdu (e.g. *Pakeezah* made in 1972 and sanskritized *Swades* made in 2004). At a detailed level, Urdu and Hindi are separate languages and deserve to be studied and treated as separate languages. This is most apparent in the official documents produced by the Indian government in Hindi and news broadcasts [8]. Noticing the growing trend of the usage of Sanskrit words in Hindi, researches of both Urdu and Hindi have started to describe them as separate languages. The commonality of the two languages is described by Matthews [8] as an *unfortunate oversimplification* of two vibrant languages. Russell [13] in his critique of Christopher King's book, "One Language Two Scripts", cites a number of examples where he shows that Hindi and Urdu are similar but different languages and sometimes the vocabulary, usage, and pronunciation can make a huge impact on understanding of the language. Russell compared language teaching books of Hindi and noted that a number of the words can easily be treated as Urdu like *akela (alone)* and *akelapan (loneliness)* but soon the difference start to appear (e.g., *adhik (lots)* in contrast to *zyada (lots)* in Urdu, *akash (sky), asman(sky)* in Urdu).

I should not use only a few examples like these to make a statement about two different languages. The translation of "How far is your house from here?" will be understood both by Hindi and Urdu speakers, but the divergence trend between these two languages continues with time (e.g., bicycle in Hindi referred to as *do chakr ghamia* whereas in Urdu it remains to be called *cycle*.) The official Hindi name for the popular game cricket is called *Gol guttam lakad battam de danadan pratiyogita*. However, this usage is completely absent from day-to-day usage in India—the borrowed word *cricket* is used instead. The following example is borrowed from Russell to explain the growing divergence. Consider the sentence in English "The eighteenth century was the period of the social, economic and political decline". The Urdu translation of the sentence is "*Atharvin sadi samaji, iqtisadi aur siyasi zaval ka daur tha"* while the Hindi equivalent is "*Atharvin sadi samajik, arthik aur rajnitik girav ki sadi thi"*. Moreover, in Hindi *"sadi"* could be replaced by *"satabdi"* and *"aur"* with *"tatha"*. Russell points out that this example alone shows that Urdu speakers cannot understand the meaning of the Hindi equivalent and vice versa. Therefore, these two languages should not be treated as the same language in all circumstances.

3.3.1 Highbrow, Middlebrow, and Lowbrow

Besides script, the most notable differences between Hindi and Urdu are found in the formalized vocabulary, grammar, and writing style. King [8] quoting Ashok Kelkar, a proponent of Hindi, describes those differences in detail as an excellent example of a social linguistics situation. Hindi and Urdu have a full range of styles. He categorized those styles as stated below:

- *Formalized highbrow* is used in academia, religious sermons, official texts, and poetry. Most language engineering resources and enabling technologies for system development are based on this style. Highbrow Hindi draws its base from Sanskrit and highbrow Urdu, throughout time, has been based on Persian and Arabic words.
- *Formalized middlebrow* is used in songs, movies, pamphlets, popular printed literature, and mass propaganda.
- *Casual middlebrow* is most widely used for daily conversations among the educated upper and middle class who are regionally based like in northern India and Hyderabad. It is used for private communication like phone conversations and letter writing. It is used by newspapers so they can be read by a large audience. This style is most receptive to borrowed words, most of them from English.
- *Casual lowbrow* is associated with what Kelkar calls the "lower class" and uneducated people. He calls it *bazaar Hindustani*. This is definitely a substandard form of style. This style is found in the slums of urban centers of large Indian cities.

The polarization of Urdu and Hindi reaches its maximum at *formalized highbrow*. Hindi draws from Sanskrit for vocabulary and promotes Vedantic and Brahmanical culture. Urdu draws from Turkish and classical Persian literature and Islamic events as references.

King [6] mentions that standard Hindi (highbrow) and standard Urdu (highbrow) have diverged more since the partition of India and Pakistan in 1947. A careful analysis of King's theory shows that it is certainly true that standard Hindi is getting more and more sanskritized, but new Urdu literature is leaning towards formalized middlebrow. Sanskritized Hindi is increasingly used by the elite in India. This movement of sanskritizing Hindi in India is illustrated by King [6] while quoting Das Gupta and Gumperz. The illustration is done by analyzing the signboards; label *a* is the official text of the signboard, label *b* is the English translation and label *c* is the causal middlebrow in Hindi. I have added label *d* as the highbrow in Urdu and label *e* is the middlebrow in Urdu.

- Signboard 1
  a. dhumrpan varjint hai
  b. smoking is prohibited
  c. cigarette pina mana hai
  d. tambakoo noshi mana hai
  e. cigarette pina mana hai

- Signboard 2
  a. Bina agya pravesh nishedh
  b. entrance prohibited without permission
  c. bina agya andar jana mana haid
  d. baghair poochey andar aana mana hai
  e. baghair poochey andar aana mana hai

Note that for signboard 2 middlebrow and highbrow is the same for Urdu.

3.3.2 Cultural differences

Although languages don't belong to a religious group, it is an undeniable fact that Urdu is the first language of the large Muslim population in India and is known to most Muslim Pakistanis as their national language. The same is true for the Hindu majority in India where most Hindus in North India prefer Hindi. I don't mean to imply that there are no Hindu scholars of Urdu, there are a number of those—Prem Chand and Gopi Chand Narang are notable examples.

The cultural preferences of speakers translate into respective languages. The date and year reference for Muslims for major

events in South Asia is the Hijri calendar (reference to Prophet Muhammad's migration from Makah to Medina). The year 2000 in C.E. is 1421 Hijri. This is evident in how different people reference the completion of the Taj Mahal for example. A Muslim cleric will refer to its completion in 1076 Hijri, but a secular Hindu will say the date is 1666 A.D. The epitaph inside the Taj Mahal refers to the Hijri date not the Gregorian calendar.

3.4 Script Differences

In this section I will explain few of the script differences between Hindi and Urdu in terms of phonemes, spoken units of a language, and graphemes, written units of a language. Hindi and Urdu have most of the phonological features of the languages of the subcontinent like retroflexion and voiceless and voiced, aspirated and unaspirated stops. The majority of the differences in Urdu and Hindi regarding the script are based on the vocabulary. Besides supporting the features of the languages of the sub-continent, the Urdu script supports the phonemes of Persian and Arabic. For example, in contrast to Hindi, Urdu has an unaspirated uvular stop /q/, labial fricative /f/, voiceless retroflex /ʂ/, velar fricative /x/, voiced dental fricative /z/, palato-alveolar voiced fricative /ʒ/, and voiced velar fricative /ɣ/. These sounds are supported by the nastaliq and naksh styles of Urdu script. Hindi has a system of making these sounds native by changing the articulation at different levels for each foreign sound. Urdu script has distinct graphemic features for retroflexion and aspiration. Urdu uses diacritic marks for retroflexion, and aspiration in Urdu is shown by *h* whereas the Devanagri script of Hindi does not treat retroflexion and aspiration as distinctive features. There are a number of other examples, but the few examples above justify the difficulty when using Hindi resources for Urdu computational processing. One of the easier tasks for language engineering is a transliteration from one language to another by using a map from one symbol to another. The work of Jawaid and Ahmed [5] shows that there are many open issues when transliterating Hindi to Urdu or vice versa. The above differences show that Hindi stemmers cannot be used for Urdu stemmers [9].

4. Quantitative Analysis of Hindi and Urdu

In this section I show a few quantitative examples where Hindi resources cannot be used for Urdu Information Access.

4.1 Named Entity Recognition

Named entity recognition (NER) is one of the important tasks in the field of Information Access. It constitutes automatically recognizing proper nouns like people names, location names, and organization names in unstructured text. There has been significant work done in English and European languages, but this task is not well-studied for the languages of South Asia.

By far the most comprehensive attempt made to study NER for South Asian and Southeast Asian languages was by the NER workshop at the International Joint Conference of Natural Language Processing in 2008 [14]. The workshop attempted to do named entity recognition in Hindi, Bengali, Telugu, Oriya, and Urdu. Among all these languages, Urdu is the only one that has Arabic script. Test and training data was provided for each language by different organizations; therefore, the quantity of the annotated data varied among different languages. A shared task was defined to find named entities in the languages chosen by the researcher. There are fifteen papers in the final proceedings of the NER workshop at IJCNLP 2008. A number of those papers tried to address all South Asian languages in general, but resorted to Hindi where the most number of resources were available. A number of papers addressed specific languages like Hindi, Bengali, and Telugu, and one paper addressed Tamil. There was not a single paper that focused on Urdu named entity recognition. In the papers that tried to address all languages, the computational model showed the lowest performance on Urdu. None of the researchers were able to use the online dictionaries and gazetteers that were available for Hindi for Urdu.

4.2 Zipf's Law on Hindi and Urdu

One of the most interesting, and maybe the only, works available that compares Urdu and Hindi quantitatively is by Martynyuk [7]. In this study Martynyuk establishes his argument based on the comparative analysis of the most frequent words used in various European languages. He then extends his work to compare Hindi and Urdu based on Zipf's Law. Zipf's law is one of the most fundamental laws used by researches when analyzing text in an automated way. On a high level, Zipf's law states that given a corpus of words, the frequency of the words is inversely proportional to the rank of the word. Although not stated clearly, the hypothesis of the experiment is that if the most frequent words show similar rank order in corpora of different languages like Hindi and Urdu, then the theory of *"same language two scripts"* has a good chance of being accepted in the larger academic community. Hindi and Urdu news text tokens were categorized using manual lemmatization. The experiment was conducted on approximately 440,000 Romanized words of Urdu and Hindi. After calculating the frequencies of the most frequent words, it was found that the top three words between Urdu and Hindi were the same words (stop words). The rank order between the words starts to divert after rank three. The top 24 ranked words are the same between the two languages, but have a different rank order than each other. After rank 24, the words start to differ in rank order and their alignment with the corresponding word from the other language. For example the Hindi word for election, *chunaav*, is at rank 25; the Urdu equivalent for election, *intikhabaat,* is found at rank 45. It is important to note that the Hindi corpus did not contain the English word *election*, but in the Urdu corpus it was used almost as many times as *intikhabaat* (*intikhabaat* is used 672 times and *election* is used 642 times.) This next example may drive home the point that the rank order of the same meaning words keep on drifting apart. The word for terrorist in Hindi, *atankvadee,* is at rank 42 in contrast to the equivalent word in Urdu, *dehshat gard*, which is at rank 182. This experiment clearly shows the Information Access features like term frequency, inverse document frequency, and stop word analysis for Hindi and Urdu will show different results [12].

4.3 Hindi Wordnet

This research is part of the larger research area doing concept searching in Urdu [10]. Wordnet is an important enabling technology for concept understanding and word sense disambiguation tasks. Hindi Wordnet [2] is an excellent source for Hindi language processing but cannot be used for Urdu. Most of the analysis of the words and the categorization of words in Hindi Wordnet is done by using highbrow Hindi. For example, the terminology used to describe parts of speech (POS) in Hindi Wordnet is completely foreign to Urdu speakers. The POS names are Sanskrit-based whereas Urdu POS words are Persian-based and Arabic-based. In Hindi the word for noun is *sangya* and in Urdu it is *ism.* The proper noun in Hindi is called *Vyakti vachak sang.* No Urdu speaker would know this unless they have studied

Hindi grammar. In order to work through these differences, one has to be familiar with both languages at almost expert levels.

5. Conclusion and Remarks

In this position paper I contend that Hindi and Urdu are two different languages, at least for computational processes. The reasoning in this paper does not suggest that Urdu and Hindi are two different languages in all aspects, but rather asserts that the current tools available for Hindi language cannot be used for Urdu language processing. In other words, in order to use Hindi resources to do Urdu computational processing, one has to know Hindi at a detailed linguistic level. The examples of the quantitative analysis confirm the emerging research from the Urdu and Hindi language researchers that there is trend of divergence between the two languages. There is a great opportunity to use casual middlebrow to bridge these languages and develop tools for multilingual Information Access. In the meantime, new resources need to be created to do research in Urdu Information Access.

6. REFERENCES

[1] Becker, D. and Riaz, K., 2002. A Study in Urdu Corpus Construction. Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics.

[2] S. Jha, D. Narayan,Pande, P., Bhattacharyya, P., 2001. A Wordnet for Hindi, International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, India

[3] Baker, P., Hardie, A., McEnery, T., and B.D. Jayaram. 2003. Corpus Data for South Asian Language Processing". Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL 2003.

[4] CRULP, 2009. http://crulp.org/software

[5] Jawaid, B., Ahmed, T., 2009. Hindi and Urdu Conversion beyond Simple Transliteration. Conference of Language and Technology.

[6] King, R., 2001. The Poisonous Potency of Script: Hindi and Urdu, International journal of the sociology of language

[7] Martynyuk, S., 2003. Statistical Approach to the Debate on Hindi and Urdu, Annual of Urdu Studies vol. 18

[8] Matthews, D., 2002. Urdu in India, Annual of Urdu Studies vol. 17

[9] Riaz, K., 2007. Challenges in Urdu Stemming, Future Directions in Information Acces.

[10] Riaz, K., 2008. Urdu Concept Searching. PIKM 08

[11] Riaz, K. Baseline for Urdu IR Evaluation, 2nd ACM iNEWS08 Workshop on improving non English Web Searching. CIKM 08

[12] Riaz, K. 2007. Stop Word Identification in Urdu. Conference of Language and Technology.

[13] Russell, R., 1996. Some Notes on Hindi and Urdu, Annual of Urdu Studies vol. 11

[14] Workshop on NER for South Asian and South East Asian Languages at IJCNLP. http://ltrc.iiit.ac.in/ner-ssea-08/ (2009)

A Patient Support System based on Crosslingual IR and Semi-supervised Learning

Hideki Isozaki, Tsutomu Hirao, Katsuhito Sudoh, Jun Suzuki,
Akinori Fujino, Hajime Tsukada, Masaaki Nagata
NTT Communication Science Laboratories
NTT Corporation
2-4 Seikacho, Sorakugun, Kyoto, Japan, 619-0237
{isozaki,hirao,sudoh,jun,a.fujino,tsukada,nagata}@cslab.kecl.ntt.co.jp

ABSTRACT

Nowadays, patients use the Web to get useful information. However, latest medical information is not available in most languages other than English. Even if patients want to learn about up-to-date treatments, they do not want to read English documents filled with technical terms. In order to mitigate this situation, we are building on a patient support system that combines crosslingual information retrieval, machine translation, and technical term extraction to provide up-to-date medical information.

Categories and Subject Descriptors

I.2.7 [**ARTIFICIAL INTELLIGENCE**]: Natural Language Processing—*Machine Translation*
; J.3 [**Computer Applications**]: LIFE AND MEDICAL SCIENCES—*Medical Information Systems*

General Terms

CLIR

Keywords

medical information system, crosslingual information retrieval

1. INTRODUCTION

Nowadays, patients use the Web to get useful information (`www.nytimes.com/2008/09/30/health/30online.html`). English-speaking patients can obtain useful medical documents from different sources such as

- Patient community sites (e.g., PatientsLikeMe) and patients' blogs.
- Reports from public medical organizations (e.g., WHO, NIH, NCI, and FDA).
- Academic documents (PubMed = abstracts of biomedical papers from the US National Library of Medicine).

SIGIR '09 Boston, MA USA

However, latest medical information is not available in most languages other than English. In order to support patients who want to learn about up-to-date treatments but do not want to read English documents filled with technical terms, we are building a patient support system.

Similar patient support services are already available on the Web. Nikkei BP's Cancer Navi (`cancernavi.nikkeibp.co.jp/`) provides cancer news in Japanese. Cancer Information Japan (`cancerinfo.tri-kobe.org/`) translates National Cancer Institute (NCI)'s Physician Data Query (PDQ) into Japanese. This site provides two versions of translation. one for experts and the other for patients. Japan Association of Medical Tranlation for Caner (`www.cancerit.jp/xoops/`) translates PubMed abstracts.

However, these services are labor intensive and costly. Our motivation is to automate them as much as possible.

Data mining and natural language processing (NLP) of biomedical literature are hot research topics (`compbio.uchsc.edu/BioNLP2009/index.shtml`), and useful BioNLP tools are available from the National Centre for Text Maining in the UK (NaCTeM, `www.nactem.ac.uk/`), Tsujii Laboratory of The University of Tokyo (`www-tsujii.is.s.u-tokyo.ac.jp/`), and so on.

The goal of our project is to analyze not only academic documents (PubMed abstracts and full papers) that are relatively uniform and written for experts, but also patients sites and reports form public/governmental organizations.

Patients sites are not always trustworthy, but they are much easier to understand and sometimes provide useful information that are not written in doctors' papers.

In future, our system will provide a stereoscopic view: medical experts' view and patients' view. The system will also integrate world-wide information (e.g., A certain medicine is available in the USA but not in Japan.) and local information (e.g., Hospital H's doctor D is good at a certain surgical operation).

Information credibility is essential in such a system. The above services explicitly disclaim endorsement and liability, and we will also follow a similar policy. However, information credibility on the web is an important research issue (`www.dl.kuis.kyoto-u.ac.jp/wicow2/`), and we will introduce some credibility handling function into our system.

2. CURRENT STATUS

The current system consists of the following modules.

- Dictionary-based crosslingual information retrieval module.

When a patient enter a Japanese term, it is translated into corresponding English terms. We use Life Science Dictionary (LSD) (`lsd.pharm.kyoto-u.ac.jp/en/`) for Japanese-to-English term translation and Indri (`www.lemurproject.org/indri/`) for document retrieval.

- Technical Term Recognizers.

  This module detects and classifies medical terms in PubMed entries, We use two technical term recognizers: a dictionary-based left-to-right longest pattern matcher and a *semi-supervised* CRF tagger [5].

- Rhetorical Parser.

  Some PubMed abstracts are tagged with `<OBJECTIVE>`, `<METHOD>`, `<RESULTS>`, and `<CONCLUSION>`. Patients are usually interested in `<RESULTS>` and `<CONCLUSION>`. They are not interested in `<METHOD>` part that describes details of experiments.

  These tags are useful in document retrieval and crosslingual multidocument summarization. For instance, we can skip translation of `<METHOD>` part for patients who do not want to read details of experiments but are eager to find effective treatments.

  In order to cover other abstracts without these rhetorical tags, we implemented a semi-supervised discource parser.

- Hierarchical Phrase-based Statistical Machine Translator[6].

  This module translates English documents into Japanese. It is trained with 20,000 medical bilingual corpus on the Web and 2,150,000 dictionary entries. We are intensively working on the improvement of this module, and will report it somethere else.

2.1 Crosslingual IR

When a Japanese patient enters "Tamiflu" in Japanese, the dictionary gives the English expression "Tamiflu." However, "Tamiflu" is rarely used in PubMed because it is a commercial product name and most medical articles use its substance name "oseltamivir phosphate" or "oseltamivir" instead.

Readers can check this fact by using MCBI's search engine (`www.ncbi.nlm.nih.gov/`). If you enter `"Tamiflu" [All Fields] NOT "oseltamivir" [All Fields]`, the query yields only 16 items. On the other hand, `"oseltamivir" [All Fields] NOT "Tamiflu" [All Fields]` yields 836 items. If you enter a simple query `"Tamiflu"`, NCBI's search engine expands it to a complex query `"oseltamivir" [MeSH Terms] OR "oseltamivir" [All Fields] OR "tamiflu" [All Fields]`, and you get 963 items.)

We extracted correspondence between commercial product names and substance names from LSD, and the system adds formal substance names for retrieval by using Indri's synonym operator. And the system tells the user about these terms. Otherwise, they will be confused by unknown terms.

A Japanese query "Tamiflu" is expanded into the following Indri query, and the system notifies users that the system also search `oseltamivir` and `oseltamivir phosphate`.

```
#syn(tamiflu oseltamivir #1(oseltamivir phosphate))
```

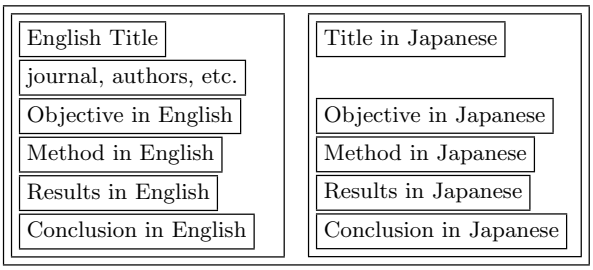


Figure 1: An entire abstract page

Our system is simiar to OReFiL [7] but OReFiL does not provide crosslingual retrieval function yet.

Patients do now want to read poorly translated technical papers. They simply want to know only conclusions: best treatments, best hospitals, best doctors, etc. From this point of view, translation has only a limited role in this system.

We also think that patients want to know trends of treatments. Therefore, the CLIR module analyzes the search engine's output. First, the number of top documents is counted for each year. The module also counts technical terms in the titles of the retrieved documents and shows frequent terms and their counts for each year.

For instance, when a patient enter "suigan" (pancreatic cancer), the system showed that `gemcitabine` appears 262 times in the titles of top 300 documents. That is, almost all paper has this medicine in the title of papers on pancreatic cancer. When a patient enter "sinkei koushu" (glioma), the system showed that `temozolomide` appeared 92 times in the titles of top 300 documents, and 48 times of them appeared in 2006 or later. It shows that treatment with this medicine became popular recently.

In this way, *document retrieval by a disease name and extraction of medicine names from titles reveal the trend of its treatments.* Patients can obtain useful information without reading poorly translated articles. This method can be regarded as a redundancy-based question answering systems [2, 4, 1] that answers a question "which medicine is used for this disease?". Since we can prepare the list of disease names, we can analyze the trend of treatments beforehand for each disease.

By clicking one of the retrieved titles, we can see the entire text of the abstract. Figure 1 shows the structure of the page. The left part of the page shows the original English abstract separated by a rhetorical parser described below. The right part shows the same abstract in Japanese. Medical terms detected by technical term recognizers in the left part are colored. When you move the cursor over them, their descriptions from different information sources pop up.

2.2 Semi-supervised learning

Semi-supervised learning methods uses We applied semi-supervised learning to rhetorical parsing and technical term extraction.

2.2.1 Rhetorical parsing

Hirohata et al.'s rhetorical parsing [3] classifies each sentence in PubMed abstracts into four classes: `OBJECTIVE, METHOD, RESULTS, and CONCLUSION`. We used their 'pure' corpus data (10,000 abstracts for training). Table 1 shows

Table 1: Accuracies of rhetorical classification

	Hirohata et al.	Our system	
	supervised	supervised	semi-supervised
per sentence	94.3%	93.9%	**95.1%**
per abstract	62.9%	61.7%	**67.7%**

that our semi-supervised learning method outperformed Hirohata et al's supervised CRF method though our implementation of the supervised CRF system have not achieved their performance yet. Here, we used 100,000 abstracts as unlabeled data for semi-supervised learning.

This rhetorical structure is now used for selecting fields on document retrieval. For instance, you can compose a complex query such that OBJECTIVE field has "pancreatic cancer" and RESULTS field has "statistically significant" and so on. We are also planning to use this classification results for multidocument summarization.

2.2.2 Medical Term Recognition

Medical term recognition can be regarded as an extension of named entity recognition popular in Natural Language Processing community. Conventional named entity recognizers usually use *supervised* learning approach such as conditional random fields (CRFs) and support vector machines (SVMs). In this approach, we do not have to write down complicated rules to detect named entities. But we have to prepare a large training data. However, their performance is almost saturated. Therefore, we employed semi-supervised approach [5], which is robust against new terms.

We trained the semi-supervised CRF by using Penn BioIE Corpus (CYP450) `bioie.ldc.upenn.edu/publications/latest_release/`. We do not know any experimental reports on this corpus yet. We have just started a preliminary experiment with this corpus. When we used 73,108 sentences for training and 8,137 sentences for test, supervised CRFs gave F-measure = 0.897 whereas semi-supervised CRFs slightly improved it to 0.905 when we used the above 100,000 PubMed abstracts as unlabeled data.

Since CYP450 covers only articles on cytochrome P450 enzymes, this tagger is not general enough. Therefore, we also built a simple left-to-right longest match tagger based on LSD.

3. CONCLUSIONS AND FUTURE WORK

We are building a CLIR system for patients who want to learn up-to-date treatments but do not want to read English documents filled with technical terms. We showed that a simple combination of CLIR and medical term recognizers give trends of treatments. As for now, the system covers only PubMed entries, which are uniform, trustworthy, and relatively easy to analyze. Future work includes analysis of patients sites and public organization sites. Introduction of patients sites will require credibility judgement.

4. REFERENCES

[1] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *TREC*, 2001.

[2] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting redundancy in question answering. In *SIGIR*, pages 358–365, 2001.

[3] K. Hirohata, N. Okazaki, S. Ananiadou, and M. Ishizuka. Identifying sections in scientific abstract using conditional random fields. In *IJCNLP*, pages 381–388, 2008.

[4] C. C. Kwok, O. Etzioni, and D. S. Weld. Scaling question answering to the web. In *WWW*, pages 150–161, 2001.

[5] J. Suzuki and H. Isozaki. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *ACL*, pages 665–673, 2008.

[6] T. Watanabe, H. Tsukada, and H. Isozaki. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of COLING-ACL-2006*, pages 777–284, 2006.

[7] Y. Yamamoto and T. Takagi. OReFiL: an online resource finder for life sciences. *BMC Bioinformatics*, 8:287, 2007.